IN THE UNITED STATES DISTRICT COURT FOR

THE DISTRICT OF MASSACHUSETTS

NUANCE COMMUNICATIONS, INC., Plaintiff and Counterclaim Defendant, v. OMILIA NATURAL LANGUAGE SOLUTIONS, LTD., Defendant and Counterclaim Plaintiff.	**Case No. 1:19-cv-11438-PBS**

DECLARATION OF CHRIS SCHMANDT IN SUPPORT OF NUANCE COMMUNICATIONS, INC.'S OPPOSITION TO DEFENDANT'S MOTION FOR PARTIAL SUMMARY JUDGMENT FOR THE INVALIDITY OF U.S. PATENT 6,999,925

I, Chris Schmandt, hereby declare and state the following:

I. INTRODUCTION AND SUMMARY OF OPINIONS

1. I was hired as an expert consultant and to provide consultant services on behalf of Nuance Communications, Inc. ("Nuance"), in connection with this litigation.

2. I have been asked to provide this declaration concerning whether claims 1 and 27 of U.S. Patent No. 6,999,925 (the '925 Patent) are invalid under the doctrine of obviousness-type double patenting ("ODP") in view of claim 6 of U.S. Patent No. 6,789,061 (the '061 Patent), and whether claim 14 of the '925 Patent is invalid under the doctrine of obviousness-type double patenting in view of claim 14 of the '061 Patent. It is my opinion that they are not, applying the legal standards as they have been explained to me. My analysis is set forth in detail below.

3. I am over the age of 18 years and I have personal knowledge of the facts recited herein and could testify competently thereto.

4. The purpose of the '925 Patent is to take an existing recognizer, and enhance it for a particular domain, such that it retains its original richness, but is also optimized for the new domain because the new speech recognizer can account for new phonetic contexts not present in the general speech recognizer. In a more general sense, ". . . one is able to derive a new, recalculated set of equivalence classes that can be considered by construction as a domain or dialect dependent refinement of the original phonetic contexts." ('925 Patent col. 8:7-10).

5. The '925 Patent describes a process to adapt a given recognizer, with an associated decision tree previously built with labelled training data. Using the existing decision network as a starting point, new domain-specific training data is run through the decision network. In one embodiment, the decision tree component of the decision network serves as the starting point, asking the same questions as before about the likelihood of achieving a better partitioning of the tree by splitting or merging nodes of the newly developed decision tree, allowing the phonetic contexts to change through this process to account for phonetic contexts that are meaningful in the training data.

6. At the end of this process, the HMMs previously associated with the leaf nodes may be re-adjusted to appropriately model the new training data. There may be more, or fewer, nodes, including leaf nodes. In short, the ’925 Patent generates a new recognizer by adapting the decision network with new training data. HMM parameters may be re-adjusted in the last stage of this process because training in the new acoustic data may suggest other speech patterns characteristic of the phonemic contexts grouped into any particular leaf node.

7. By contrast, the purpose of the ’061 Patent is to take an existing recognizer and produce a new, smaller recognizer for a particular domain, such that the new recognizer requires less processing power or a smaller memory footprint using the existing phonetic contexts of the general speech recognizer.

8. That is, the ’061 Patent adapts a given recognizer to a new domain only by deleting some of the HMM states and associated PDFs (especially if an HMM is associated with PDFs in the form of Gaussian Mixture Models (GMMs) of multiple underlying base PDFs). This is accomplished by running training data through the recognizer, and noting which HMMs are not activated, or for those that are, any of the mixed PDFs that are not relevant (don’t contribute much to the probability of matching an input acoustic frame). Because the only operation on the HMMs is deletion (of whole HMMs or only some of their associated PDFs) the new recognizer is smaller, a subset of the previous recognizer.

9. The heart of Omilia’s argument seems to be that the HMM states which may be deleted in the ’061 Patent are the same as the nodes of the decision tree which may be deleted (among other operations) in the ’925 Patent, and that therefore the two patents are simply stating the same invention in different terms. This is clearly incorrect.

10. The nodes of the decision tree do not correspond to HMM states, although they can be used in the process of generating HMM states. HMM states model production of speech sounds,

based on the probabilities of each HMM producing a given set of feature vectors.[1] Decision tree nodes are binary questions about phonemic context that are used to separate training data acoustic frames, the final set of which will be used to construct the HMMs.

11. HMMs have associated PDFs (related to matching acoustic features), as stated in the preamble of each independent claim. There are no PDFs associated with decision tree nodes.

12. Contrary to Dr. Cohen's assertions, the decision trees used to generate the recognizer by portioning the training speech data are not used to recognize speech. Recognition involves capturing speech, converting it to acoustic frames, and then calculating which HMM sequence was most likely to have produced that series of frames. Decision trees can be used during training of the acoustic model by sorting out labeled speech recognition data, or to assist in the sequencing of phonemes that is recognized by the acoustic model when deployed.

13. When acoustic features are passed through the decision tree as part of the training process, they must be labelled. The question at each node of the decision tree asks which phoneme this frame is from, and in which phonemic context; the decision does not look at the feature data to make this decision, only the labels. This is because decision tree process clusters sets of training frames based on phonemic context; only then are the acoustic contents of the frames analyzed to determine an appropriate HMM and PDF(s). Similarly, when decision trees are used in improving the performance of a speech recognizer, the trees refine sequences of phonemes that were recognized by the acoustic models.

14. Although the '925 and '061 Patents both teach techniques to produce a new (second) speech recognizer, which is improved according to some (different) metrics, and even though both techniques start with a (first) general purpose speech recognizer and make use of specialized training data, they use completely different and disjoint techniques, neither of which is a "species" or "instance" of the other.

[1] Within the recognizer, speech is converted to feature vectors, a computationally concise way to mathematically represent the speech sounds in spoken audio.

II. QUALIFICATIONS

15. I received my BS degree in Computer Science in 1979 and my MS degree in Visual Studies (specifically interactive computer graphics) in 1980, both from MIT. Following my studies, I spent my entire career at MIT, first as a research scientist at the Architecture Machine Group, then as a Principal Research Scientist at the Media Laboratory, of which I was a founding member in 1985. In the Media Laboratory I served as a faculty member and research supervisor, and directed a research group with student members at all levels. This group, Speech Research, focused on applications and user interfaces for speech technologies, with an emphasis on telecommunication systems.

16. In these positions I authored research proposals, performed and supervised said research, and taught classes and supervised theses at all levels. I also served on the Departmental Intellectual Property committee and chaired it for about five years. I supervised all undergraduate research in the Laboratory since its inception, and for the last half dozen years of my career supervised thesis proposal and general examination approval for all graduate students in the Laboratory.

17. I regularly published papers, of which I have approximately 80, usually with my students as co-authors. I also wrote a book, Conversational Computer Systems, in 1996; this book discussed technologies and applications for speech recognition and synthesis, digital audio recording, voice telecommunications systems, and interactive voice response. For fifteen years I taught a graduate level course based on this text. I am also a named inventor on a number of US Patents.

18. Speech technologies were the core of my research for the first two decades of my career at MIT. I became well known for my early Put That There system, which was one of the very first multimodal computer interfaces (1979) employing speech recognition and also a pioneering conversational voice system. Much of my work focused on applications of speech technologies and user interface techniques, and as such included work in language models, parsers, and dialog design. With my students I built one of the first unified messaging systems, Phone

Slave, in 1983; the system integrated voice and text messages, with telephone-based speech and desktop oriented graphical user interfaces. I also built the first real time spoken driving directions navigation system in 1988; as with the systems that became commonplace a decade later, Back Seat Driver could determine optimal routes and instruct a driver to follow them, including detecting and correcting navigation errors. I applied speech recognition to user interface design of hand-held and wearable computing and communication systems, such as Voice Notes and Nomadic Radio, which were early hand-held and wearable computers equipped with speech recognition input.

19. For my pioneering work in voice user interface design, I was awarded a position in 2007 into the Association for Computing Machinery's CHI (Computer Human Interaction) Academy, an honor given to a handful of researchers each year.

20. I retired from MIT at the beginning of 2019.

21. Attached hereto as **Exhibit A** is a copy of my curriculum vitae, which provides additional detail about my background and qualifications.

III. MATERIALS CONSIDERED

22. To prepare this declaration, I have considered the '061 and '925 Patents, their prosecution histories, the declaration of Jordan Cohen in support of Omilia's motion for partial summary judgment, and the materials cited in his declaration. I have also considered the materials that are discussed in the body of this declaration. I have also considered certain legal standards, discussed later in this declaration, that I understand apply in patent litigation.

23. I have prepared this declaration based upon the information and materials that are currently available. In the event that additional information becomes available or is provided to me, then I may change, supplement, or amend my analysis and opinions based on any newly available information.

IV. LEGAL STANDARDS

24. To be clear, I am not a lawyer or an expert in patent law. I have been informed that it is ultimately the Court that will decide and apply the legal standards related to Omilia's motion. Following is my understanding of the legal standards that the Court will apply in its analysis.

A. Legal Standards Concerning Obviousness-Type Double Patenting.

25. In connection with this declaration, I have been asked to assume the following about obviousness-type double patenting.

26. I understand that the doctrine of double patenting is intended to prevent a patentee from obtaining a time-wise extension of a patent for the same invention or an obvious modification thereof. In other words, I understand the double patenting doctrine to be aimed at situations where two patents from the same inventor, with different expiration dates, claim the same invention or an obvious variant thereof. Although I understand that there are two varieties of double patenting, "same invention" (covering identical subject matter) and "obviousness-type" (where a later-expiring patent is an obvious variant of an earlier one), I understand that Omilia's present motion only relates to obviousness-type double patenting ("ODP").

27. I understand that analyzing claims for obvious-type double patenting requires comparing claims in an earlier patent to claims in a later patent. Thus, the starting point in such an analysis is the reference claims of the earlier patent. There are two steps in comparing the claims of the two patents. First, the claims must be construed, and any differences identified. Second, the differences must be considered to determine whether they render the claims patentably distinct. A later patent claim is not patentably distinct from an earlier patent claim if the later claim is obvious over, or anticipated by, the earlier claim. For this second step, the inquiry is analogous to the obviousness analysis under 35 U.S.C. § 103, but is limited to comparing the claims of the two patents.

28. While the analysis of the differences is to be considered on a term-by-term basis, it is also necessary to consider the claims as a whole in determining whether the later claim is patentably distinct from the earlier claim.

29. Finally, to reiterate, I understand that the focus of the ODP doctrine rests on preventing a patentee from *claiming* an obvious variant of what it has previously claimed, not what it has previously *disclosed*. Thus, it is the claims, not the specification, that define an invention. And it is the claims, not the specification, that are compared when assessing double patenting.

B. Legal Standards Concerning Claim Construction.

30. As I noted above, it is my understanding that the first step in an ODP analysis requires that the claims be construed. Accordingly, I have been asked to assume the following about patent claim construction.

31. The scope and meaning of a patent claim is determined based on what the claim would mean to a person of ordinary skill in the art ("POSITA"). This determination is made as of the time of the invention and based on intrinsic and extrinsic evidence. In my discussion below, I address what I believe to be the level of skill required for a POSITA for the '061 and '925 Patents.

32. A patent's claim language and specification are intrinsic evidence. One should analyze the intrinsic evidence from the perspective of a POSITA to understand a claim term. A patentee may define or ascribe a particular meaning to a claim term in the specification. However, a POSITA should not read limitations from the specification into a claim if the limitation is not already found in the claim language.

33. The prosecution history of a patent is also intrinsic evidence. This is the back and forth communications between the patent applicant and the U.S. Patent and Trademark Office ("PTO"), which creates a record of the patent's prosecution. The prosecution history may indicate how the claims should be interpreted, how the inventor (or the PTO) understood the patent, and differences between the claims and the prior art. My understanding is that amendments to the claim language or arguments made by the applicant during prosecution can be indicative of or limit the claim scope or meaning in the patent that issues from that application, as well as in related patents.

34. To determine the meaning of a claim term and claim scope, one should read and analyze a claim term in the context of the entire patent in which the term appears, including its specification and prosecution history, and not just in the context of a particular claim. Also, one should consider the education level, knowledge, and experience of a POSITA when determining the meaning of a claim term.

35. Extrinsic evidence can also be considered when determining the meaning of a patent claim, but it is not as important as, and cannot supersede the meaning according to, the intrinsic evidence. Extrinsic evidence can include technical dictionaries and expert testimony.

C. Definition of Person of Skill in the Art.

36. I understand claims are interpreted from the perspective of a POSITA at the time of the invention. To determine the characteristics of a POSITA for the Nuance Patents at the times of the inventions, I have considered the state-of-the-art of speech recognition systems in the relevant time period for each of the Nuance Patents. This includes considering the problems encountered in the art at that time. I also have considered the education and experience of persons who were working in the field at that time, including my own experience and knowledge at that time and others who I worked with on these issues and problems. Further, I have considered the experience, education, and knowledge of persons in industry and academia at that time who were working in the aforementioned fields.

37. The ’925 Patent was filed on November 13, 2001 and claims priority from EP application 00124795, filed on November 14, 2000. For purposes of this declaration, I consider the time of the invention to be the November 2000 timeframe when the EP application was filed. It is my opinion that a POSITA for the ’925 Patent would have at least a bachelor’s degree in computer science, electrical engineering, or another scientific field, plus two years work experience in the field of speech recognition, or education and work experience equivalent thereto.

38. I understand that Dr. Cohen has opined that a POSITA as of November 14, 2000 “would have at least a Bachelor’s of Science degree in electrical engineering and or computer

science, with preferably a Master's degree, and 2-3 years of experience working large vocabulary speech recognition systems." Cohen Decl. ¶ 25.

39. In my opinion, the differences between the level of skill in the art I have identified and that identified by Dr. Cohen are not substantial and do not make a difference in the opinions I have provided herein. As stated above and in my CV, under either articulation I have at least the qualifications of a POSITA – both today and at the time of the invention of the '925 Patent.

V. BACKGROUND TECHNOLOGY

40. The goal of speech recognition is to turn human spoken language into text. Rather than recognizing whole words, the recognizer detects sub-word units, which constitute a speech "alphabet" and hence can spell out any word. The unique sounds which make up the words of any language are called phonemes, and a recognizer must understand the stream of phonemes in our speech in order to find the words. We teach the recognizer how to identify phonemes in speech by *training* it with lots of examples of real speech making those sounds. As part of this process we must decide exactly what training data to use, and then how to build the recognizer's internal representation, or model, of that collective training data so it can find the same sounds in unknown speech input during recognition.

41. In English there are a more phonemes (linguists usually say around 42) than there are letters, because sometimes a letter can make multiple sounds. It is natural for speech recognizers to use phonemes, because phonemes are the *sounds* of a language; but recognizers must deal with the problem that the same phoneme is not always spoken the same way. Consider the "t" in "cat" versus "bottle"; they sound rather different. The sounds (phonemes, including possible silence at the ends of a word) influence the way the phoneme is spoken. A speech recognizer includes an "acoustic model" which analyzes the actual sounds of speech to map them to higher level linguistic constructs, such as these phonemes. One way this overall acoustic model is obtained is to build a specific acoustic model for how each phoneme sounds. The implication

of differing pronunciations is that multiple such specific acoustic models may be required for a single phoneme.

42. To generalize, pronunciation of phonemes depends on phonemic context (the neighboring phonemes to the left and right—speech recognition often refers to sequences of three phonemes as a triphone). One could build an acoustic model to detect each phoneme based on its sound in the context of each other phoneme preceding or following, but this would get very large, about 75,000 such models. But not all these distinct phonemic contexts really sound different; "t" in "hit" sounds like "t" in "cat" and the "t's" in "little" and "bottle" also sound the same. If we could combine the ones that sound similar, we could save a lot of memory and computation.

43. The challenge is to determine a minimal set of phonetic contexts which captures when a phoneme is invariant (sounds the same in some number of different contexts) as well as in which contexts the same phoneme will sound different. We can utilize a machine learning construct called a **decision tree** in speech recognition to determine which phonemic contexts are actually different enough to warrant their own models, and in the process use them to sort acoustic training data into buckets which can later be turned into acoustic models.[2]

44. Specifically, in the context of generating a recognizer, a decision tree is used to classify *labelled* acoustic data *frames*. While listening to speech, a recognizer samples the speech at a regular interval, such as every 10 milliseconds, and extracts certain audible features to represent what it sounds like; this basic unit of speech sound is a *frame*. Training data consists of audio recordings with time-aligned transcriptions, i.e. each acoustic frame *labelled* with the phoneme in which it occurred. Because the audio frame is labelled, we know which phoneme was being spoken when it was produced. The time alignment is necessary because most phonemes are long enough to span multiple frames and labelling will allow the frames to be sorted by phoneme.

[2] There are other ways a decision tree can be employed in speech recognition. Here I limit my discussion to context framed by Omilia in equating decision tree to recognizer states, i.e. to partition phonetic context data.

So the labelled training data consists of a large number of speech frames, and we know which phoneme was being spoken for each.

45. Each acoustic frame is passed through the decision tree, where it is subject to a series of yes/no (binary) questions based on its phonemic context (typically the phonemes before and after the phoneme with which the frame has been labelled). For example, after asking "Is it followed by an L?" the next question might be "Is it preceded by an S?" Starting from its root, then, the decision tree has many branches, coming out two at a time (for "yes" and "no") from junctions called *nodes* (a term common in discussion of graphs and networks). As the frame is subject to questions at each node of the tree, it propagates "downward" until it ends up at a *leaf* node. Initially, the decision tree may be *generated* on the fly depending on the distributions of the labelled training frames. This process strives to collect similar phonemic contexts (again, a phoneme in the context of the preceding and following phoneme) together, so we don't have to build a distinct acoustic model for every 75,000 or so phonemic contexts. At the same time, it must distinguish those contexts which sound different or else recognition will fail. Split and merge algorithms handle generating the tree, adding and deleting nodes (questions) as needed to obtain good distribution of the training data. Many of the contexts may be "equivalent" in terms of pronunciation, and hence require only a single model.

46. During this time, the recognizer is not performing recognition, but rather grouping the acoustic frames of the training data, saving each at one of the leaf nodes. Note that each of the frames that collect at the leaf nodes are *not identical*. Speech is too variable. But after this process they should all be somewhat similar, or at least cluster into similar groups. And an important function of the decision tree is to ensure that each of the final leaf node groups has a similar amount of training data, to adequately represent the entire language, but trying to avoid contexts for which there is little training data, i.e. phonetic contexts that are uncommon in the language. This is referred to as *partitioning* the training data.

47. Once the training data has been collected and grouped using the decision tree, the next step is to build the acoustic model, specific to each phoneme or, more precisely, specific to

each leaf node of the decision tree; for this discussion the model is based on **Hidden Markov Models** (HMMs). An HMM is a series of *states,* with transition *probabilities* between states, wherein at each interval of time whichever state is currently occupied can be thought of as "emitting" a particular characteristic acoustic frame. The HMM is not itself actual speech data, but rather a computational model which executes over time and generates hypothetical speech data (in a probabilistic manner).

48. After the decision tree has been generated and the acoustic training data (frames) passed through it, each leaf node of the tree has an associated collection of frames. These frames consist of actual speech data; because the training data also has a time aligned transcription, the decision tree "questions" group the training frames into phonetically significant clusters. Then an algorithm (e.g. Baum Welch) is run on that acoustic data to build an HMM which is likely to produce frames similar to those collected. An associated probability density function (PDF) relates how likely the HMM is to produce any of the frames which have been used to build it. Basically, the HMM should be likely to produce frames which are very similar to those from which it was built.

49. Each Hidden Markov Model state is associated with a particular phoneme in context cluster. And the PDF describes what "sound" (more precisely, acoustic features, defined probabilistically) that HMM state can make. The model is called "Hidden" because when we are recognizing speech, we don't know which state the HMM should actually occupy at any time because we don't yet know what is being spoken, that is the problem to solve. Rather the HMM-based recognizer will determine which states would be most likely to sound like what was actually spoken; we then assume that was what the person was saying.

50. Because we know the path through the decision tree that each acoustic frame took, and the paths are based on questions about phonemic context, we know the phoneme(s) in the context which each HMM models. This set of HMMs and associated PDFs relating HMMs to acoustic evidence constitutes a type of an **acoustic model**.

51. In performing recognition using HMMs, a series of acoustic frames is captured from speech by digitizing and converting it to the same type of acoustic frames that were used to train the recognizer. The problem is then to determine, for a given set of captured acoustic frames, which sets of HMMs, traversed in some sequence stepwise over time, would be most likely to have produced the observed spoken sequence. This is a probabilistic process – different paths could be taken through the HMMs, and some are more likely or less likely to have produced that sequence. An algorithm such as Viterbi performs this task in a well-known manner.

52. Once we have an estimation of which HMMs were traversed, we can assemble a series of phonemic contexts which those HMMs represent, and get a (likely imperfect) phonemic transcript of the speech of possible outcomes with varying probabilities of phonetic transcripts the person may have said.

53. The description of the PDFs above was a bit simplified. Note that the leaf nodes of the decision tree, which result in distinct HMMs, depend on only the phonemic context. There is some natural variability in the way distinct individuals pronounce the sounds in speaking those phonemes, but all the acoustic contexts end up being modelled by a single HMM. The pronunciations are likely to group into clusters, which we hear as the people's speaking styles. When the pronunciations cluster, this will be seen as the collection of acoustic frames to be modeled clustering in a corresponding manner. Instead of having a single PDF associated with the HMM which is built based on the acoustic data, it is often seen useful to instead mix multiple PDFs, with one PDF for each acoustic cluster. This is referred to as a Gaussian Mixture Model. When used, then, there may be multiple PDFs which associate observable acoustic features with underlying HMMs which might have generated them. If this technique is used, an HMM may have multiple associated PDFs.

VI. THE '061 PATENT

54. The primary purpose of the invention is to minimize the resources required of a speech recognizer, so it can run on a less powerful computational platform and/or with a smaller

memory footprint. *See* in general col. 1:66 – 2:35. The patent describes a way to determine which parts of the acoustic model of an initial, general purpose, recognizer are required for adequate performance on a specific task or domain, and then remove the parts which do not contribute much to recognition performance, to obtain a smaller recognizer, which may be important in circumstances where the general purpose recognizer would require too many resources (*e.g.,* memory and/or computing power).

55. This is referred to as "squeezing" and "pruning" the recognizer, *i.e.* removing, or subsetting, some portions of the acoustic model which are less appropriate for a specific and limited application, as opposed to the more general purpose speech recognizer. Specifically, this is done by starting with the set of states of the general recognizer, then choosing a subset of HMM states – squeezing - and pronunciation probability density functions (PDFs) – pruning - associated with some of those states (col. 2:15-25), based on which of those HMM states and PDFs are most likely to have produced each frame of the application-specific speech training data (col. 7:3-8). This process does not have the ability to change the phonetic contexts of the general purpose speech recognizer (and in fact it has no access to the phonetic context decision tree). It can only delete some of the *states,* which had been previously derived by sorting acoustic training data using the decision tree based on their phonetic contexts.

56. To squeeze HMM states, the training data is "Viterbi aligned" with the set of HMM states. Viterbi alignment is a process to determine which of the states would most likely produce a correct transcription, on a frame-by-fame basis. Those states which appear frequently after alignment with the training data are more important to correct recognition and must be retained. Those states which are seen infrequently may be subject to squeezing (col. 6:8-33).

57. Similarly, the sets of PDFs can be pruned by comparing the acoustic outputs that each surviving state produces versus each matched acoustic frame. Since the GMMs (Gaussian Mixture Models) describe a mixture of base output probability PDFs, it is possible to determine how similar the various PDFs in the mixture are to the actually recorded acoustic features, on a frame by frame basis. If the training data contains speech which maps well to only one of the

acoustic outputs, the other PDFs may be pruned (col. 6:34-51). Put another way, if an HMM state can, via its acoustic realization PDFs, cover several ways that particular phonemic context may be spoken, but only one of those ways of speaking is encountered in the training data, then the remaining (non-encountered) ways of speaking can be removed.

VII. THE ’925 PATENT

A. ’925 Patent’s Invention.

58. I understand that Dr. Karen Livescu previously submitted a declaration in connection with claim construction proceedings in this litigation, and that in her declaration she provided the following summary of the ’925 Patent. I agree with Dr. Livescu’s summary.

59. The ’925 Patent teaches an improved method for efficiently refining a speech recognizer to better handle particular domains (such as languages, dialects, task areas, or user-specific information, *see* col. 6:3-8, 9:6-23). At the time of the claimed invention of the ’925 Patent, general purpose, large vocabulary, continuous speech (the way humans speak when talking normally, and a more difficult type of speech for recognizers to process as compared to words spoken with pauses between each word) recognizers were known in the art and were in general usage. However, while such speech recognizers provide good overall speech recognition quality, when a use-case centers on a specific domain – which may differ from the domain of the general-purpose speech recognizer in terms of its vocabulary, dialect, acoustic conditions, and so on – the speech recognition accuracy may not be as high as would be possible with a more specialized speech recognizer.

60. While additional domains may be similar or related to the general-purpose domain, there may be particular phonetic contexts—particular ways that words are pronounced, or that words or sub-word units are used in sequence together—that are characteristic of the domain. In such cases, taking domain-specific data into consideration can improve the accuracy of a speech recognizer for a particular context of use. But using the invention of the ’925 Patent, there is not a need to completely train the new speech recognizer from scratch, particularly if there are parts

(even, perhaps, substantial parts) of the original speech recognizer that can be used in combination with a (smaller) set of specialized data to adapt the recognizer to the new domain.

61. Prior art methods of providing domain-specific speech recognizers tended to entail either merely "select[ing] a domain specific subset from the phonetic context inventory" of the general-purpose recognizer, or extensively-retraining a speech recognizer, based on a large, specialized dataset. *See* generally col. 1:15 - 2:17, quotation from col. 2:1-3. By contrast, the inventions claimed in the '925 Patent start with a general purpose, large vocabulary speech recognizer, and by adding only a modest amount of additional specialized training data ("adaptation data," *see* col. 2:56) to the initial dataset, are able to efficiently provide a new, specialized speech recognizer that includes the new training data and new phonetic contexts. Because the collection of training data and subsequent training of a speech recognizer is expensive and time consuming, adapting a general purpose speech recognizer to a specific domain greatly reduces development costs and time to market.

62. The technique disclosed in the '925 Patent is to start with the original decision network and efficiently recalculate it via the use of a smaller amount of domain-specific training data. The data is processed through the decision network in the same way as the original training data was processed to produce the original recognizer, and the new data is allowed to modify the decision tree in the process. The split-and-merge methods described for both the original and new adaptation data control the decision tree's topology. In the process of keeping a balance while partitioning the training data – separating it into the various leaf nodes of the tree to provide audio frame material to create the acoustic model – nodes may be added or removed at any level in the decision tree. Once all the adaptation data has been processed and a new phonemic decision tree thereby created, the acoustic model can be generated.

B. The '925 Patent Differentiates Itself from the European Patent Application from Which the '061 Patent Claims Priority.

63. The inventors of the '925 Patent acknowledge and distinguish that patent from the techniques of the '061 Patent. Repeatedly throughout the '925 Patent, the specification discusses

and distinguishes European patent application EP 99116684.4, which the ’061 Patent indicates on its face is the application from which the ’061 Patent claims priority.

64. Starting at col. 6:32, the inventors describe the training procedure of a speech recognizer as a two stage process, “1.) the determination of *relevant* acoustic contexts and 2.) the estimation of acoustic model parameters.” (Col. 6:34-36 (emphasis added). They then describe how some existing approaches (MAP, MLLR) focus exclusively on re-estimation of the acoustic model parameters, “Importantly, these approaches leave the phonetic contexts unchanged.” (Col. 6:44). At this point, the inventors have articulated two distinct approaches, and the ’925 Patent’s invention will be seen (in the rest of the specification) to correspond to the first. They point out that the second set of techniques have heretofore been mostly used to adapt speaker independent recognizers to a particular user (col. 6:47-49).

65. The inventors go on (col. 6:50) to discuss additional techniques of modifying acoustic models to achieve better domain specific speaker independent recognition. Here they cite to EP 99116684.4 (from which the ’061 Patent claims priority) describing it as “selecting a subset of probability density functions (PDFs) being distinctive for the domain.” (Col. 6:63-64).

66. They specifically distinguish the ’925 Patent invention from the EP invention, characterizing the two approaches as “orthogonal[] to” each other, because the ’925 Patent invention “focuses on re-estimation of phonetic contexts or – in other words – the adaptation of the recognizer’s sub-word inventory to a special domain.” Col. 6:66 - 7:2. This phonetic context is what was called out in part 1) (above) of the description of the bifurcation of two major approaches at col. 6:32. As the inventors state here, the ’925 Patent recalculates the phonetic context, while the ’061 Patent has nothing to do with it.

67. Column 7 goes on to explain clearly how the invention of the ’925 Patent is different: “Whereas in any speaker adaptation algorithm, as well as in the above mentioned documents of V. *Fischer et al.,* the phonetic contexts once estimated by the training procedure are fixed, the present invention utilizes a small amount of upfront training data for the domain specific insertion, deletion, or adaptation of phones in their respective context. Thus re-estimation of the

phonetic contexts refers to a (complete) recalculation of the decision network and its corresponding phonetic contexts based on the general speech recognizer decision network. This is considerably different from just 'selecting' a subset of the general speech recognizer decision network and phonetic contexts or simply 'enhancing' the decision network by making a leaf node an interior node by attaching a new sub-tree with new leaf nodes and further phonetic contexts." (Col. 7:2-17).

68. To be clear, the inventors themselves describe above how the '925 and '061 Patents differ. In no uncertain terms, they state that the '925 Patent's invention "is considerably different" from the '061 Patent, and that the two patents describe two distinct approaches to solving two different problems in computerized speech recognition. (Col. 7:12).

VIII. OBVIOUSNESS-DOUBLE TYPE PATENTING ANALYSIS

A. Claim 1 of the '925 Patent Compared With Claim 6 of the '061 Patent.

69. As noted above, ODP is determined on a claim-by-claim basis. It involves a two-step process. First, the claim terms are construed and compared, so any differences may be identified. Second, those differences are evaluated to determine whether the later claim (*e.g.*, '925 Patent claim 1) is an obvious variant of the earlier ("reference") claim (*e.g.,* '061 Patent claim 6). This process involves a comparison of both the individual claim terms, and the claim as a whole.

70. I understand that Omilia has argued that '061 Patent claim 6 is the reference claim. I note that claim 6 depends from claim 5, which in turn depends from claim 4, which in turn depends from claim 2, which in turn depends from claim 1. Accordingly, what Omilia refers to as "claim 6" includes all the limitations of claims 1, 2, 4, 5 and 6.

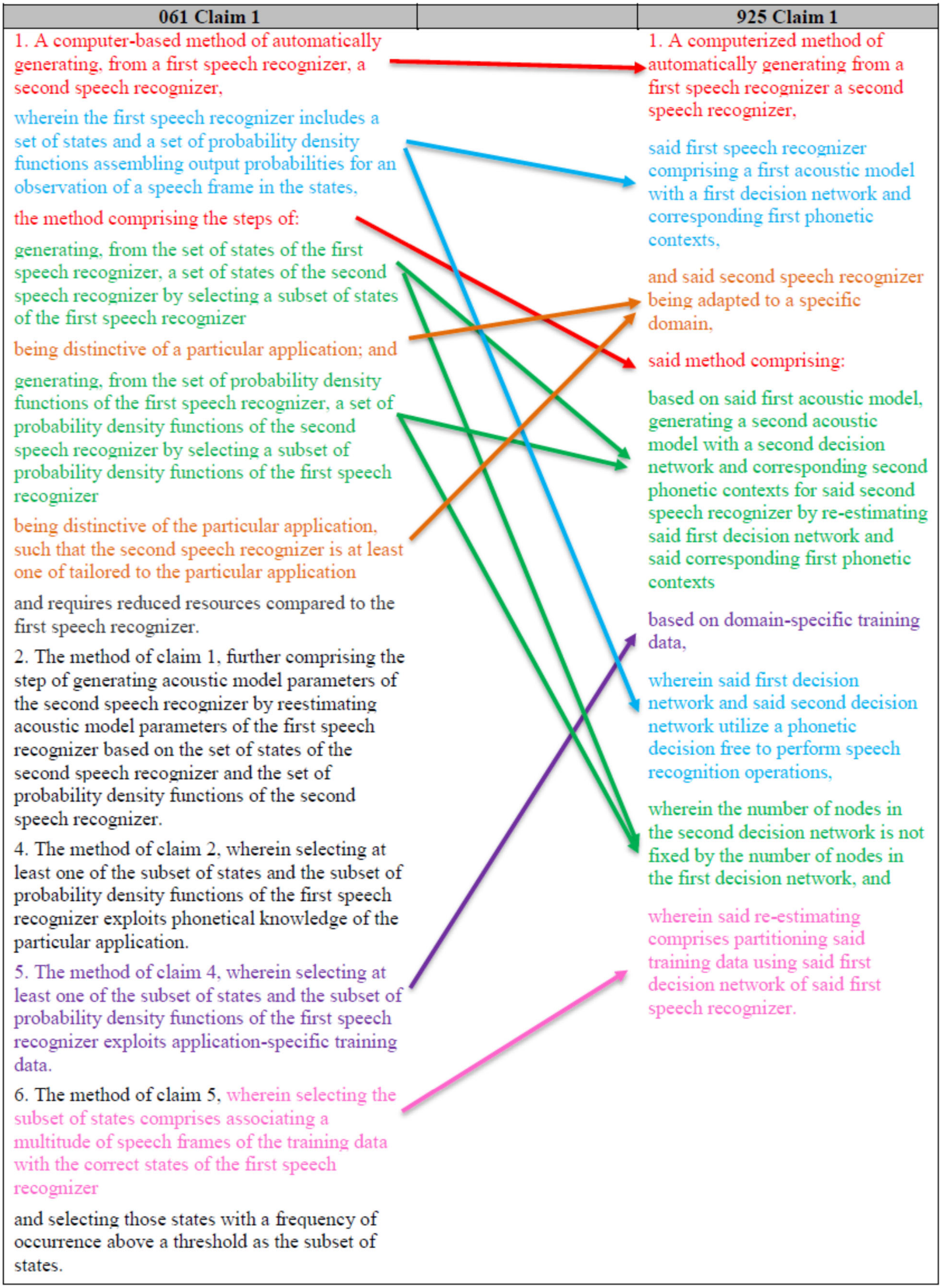
71. I have reviewed Dr. Cohen's declaration. Although he identifies a number of terms in '925 Patent claim 1 that he opines are "the same as" (or some variant of that wording) differently-worded terms in '061 Patent claim 6, he does not actually proffer constructions of any of those terms from either patent. Nor does he employ the claim construction methodology that has been explained to me and set forth above. Nonetheless as I understand Dr. Cohen's

declaration, it appears to be his opinion that there are *no* differences between these two claims. Consequently, I have been unable to locate in his declaration any analysis of whether ’925 Patent claim 1’s differences are merely obvious variants of terms found in ’061 Patent claim 1. Nor have I located any analysis in his declaration of the differences between the claims as a whole.

72. In his declaration, at paragraph 40, Dr. Cohen provides a color-coded side-by-side comparison of the two claims. Notwithstanding his color-coding scheme, Dr. Cohen’s term-by-term comparisons are not readily apparent from this diagram. Nor is it readily apparent from either his diagram or his ensuing analysis that he has “used” several claim terms multiple times.

73. Accordingly, I have prepared a modified version of his chart, with lines connecting the terms according to Dr. Cohen’s argument.[3]

[3] An additional difference is that I have placed the ’061 Patent, the older, reference, patent on the left, and the ’925 Patent on the right.



061 Claim 1		925 Claim 1
1. A computer-based method of automatically generating, from a first speech recognizer, a second speech recognizer,		1. A computerized method of automatically generating from a first speech recognizer a second speech recognizer,
wherein the first speech recognizer includes a set of states and a set of probability density functions assembling output probabilities for an observation of a speech frame in the states,		said first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts,
the method comprising the steps of:		
generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer		and said second speech recognizer being adapted to a specific domain,
being distinctive of a particular application; and		said method comprising:
generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer		based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts
being distinctive of the particular application, such that the second speech recognizer is at least one of tailored to the particular application		
and requires reduced resources compared to the first speech recognizer.		based on domain-specific training data,
2. The method of claim 1, further comprising the step of generating acoustic model parameters of the second speech recognizer by reestimating acoustic model parameters of the first speech recognizer based on the set of states of the second speech recognizer and the set of probability density functions of the second speech recognizer.		wherein said first decision network and said second decision network utilize a phonetic decision free to perform speech recognition operations,
4. The method of claim 2, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits phonetical knowledge of the particular application.		wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network, and
5. The method of claim 4, wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application-specific training data.		wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer.
6. The method of claim 5, wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer		
and selecting those states with a frequency of occurrence above a threshold as the subset of states.		

74. I disagree with Dr. Cohen's analysis.

75. I have compared claim 6 of the '061 Patent and claim 1 of the '925 Patent to identify the differences between the claims, and to determine whether the differences render claim 1 of the '925 Patent patentably distinct from the '061 Patent. It is my opinion that claim 1 of the '925 Patent is directed to a distinctly different method of developing a second speech recognizer from claim 6 of the '061 patent. That is, the claim limitations are different in kind and claim 1 of the '925 Patent is not an obvious variant of claim 6 of the '061 Patent.

B. Step One: Differences Between Claim 6 of the '061 Patent and Claim 1 of the '925 Patent.

76. One might mistakenly conclude that the '061 and '925 Patents cover the same concepts or inventions based on first glance similarities, such as the references to generating a second speech recognizer from a first speech recognizer and use of a focused set of training data. However, such similarities are in fact superficial and only establish the context within which the two patents' very different aims and technological improvements are employed. There are other similarities as well, none of which undermine the difference between the two patents' inventions. Both patents operate on speech recognizers which are described as using Hidden Markov Models. But as each patent points out, Hidden Markov Models are a commonly used speech recognition technique (and were particularly popular at the time of the inventions). This is supported by Omilia's expert. Both patents reference IBM's ViaVoice recognizer as a test platform, but as the work for both patents was performed at IBM, this should be no surprise. Both patents have similar descriptions of the state of the art, but this is simply a statement of fact and the inventors' knowledge at the time.

77. Based on a plain reading of claim 1 of the '925 Patent, it is clear that the key claim terms are not mentioned in claim 6 of the '061 Patent. Dr. Cohen argues, however, that any differences in the claims are non-substantive wording variations such that even though there is variation in the claim limitations, the limitations should be excluded from step two of the obviousness-type double patenting analysis. Specifically, he argues that the claim limitations of

claim 1 of the ’925 Patent are broader claim terms (genus) that incorporate claim elements of claim 6 of the ’061 Patent (species). I disagree with Dr. Cohen’s analysis, and in particular with his analysis of several different claim terms that he concludes are purely “semantic.”

78. Below I have identified the clauses and terms of the ’925 Patent of claim 1 that I believe are different from reference claim 6 of the ’061 Patent, and have explained why a POSITA would conclude the claims are different, and why I disagree with Dr. Cohen’s arguments. Furthermore, in view of the requirement that the claims must also be considered as a whole, I have considered the clauses and terms of claim 6 of the ’061 Patent that do not appear, and thus, are different from, the claims of the ’925 Patent.

79. I further note that because Dr. Cohen has not proffered constructions for any of the claim terms identified in his analysis, there are no claim constructions for me to respond to in connection with step one of the ODP analysis. Instead, I have based my analysis below on my understanding of how the various claim terms would be understood by a POSITA in the context of the claims, and I explain the differences in that light.

1. “said first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts” (blue) [Cohen Decl. § VIII.C.2]

Claim 6 of ’061 Patent	Claim 1 of ’925 Patent
wherein the first speech recognizer includes a set of states and a set of probability density functions assembling output probabilities for an observation of a speech frame in the states,	said first speech recognizer comprising a first acoustic model with a first decision network and corresponding first phonetic contexts

80. Claim 1 of the ’925 Patent claims a speech recognizer that comprises a “first acoustic model with a first decision network and corresponding first phonetic contexts.” Claim 6 of the ’061 Patent claims a speech recognizer that includes “a set of states and a set of probability functions assembling output probabilities for an observation of a speech frame in the states.” The decision network of the ’925 Patent operates at the level of the phonetic structure of a language: how phonemes combine and their relative frequencies for partitioning training data. The ’061

Patent is directed toward two specific aspects of a computational model for analyzing the acoustic, or sound, structure of speech.

81. A POSITA would understand that these claims are directed to different components of a speech recognizer. The wording of the ’925 Patent claim 1 is taken directly from the specification (col. 2:34-36). As I described above, a decision network (in the context of an acoustic model) can be implemented as a tree-like structure with nodes and branches. In the preferred embodiment in the specification, each node constitutes a question about phonemic context, for the sake of sorting acoustic data (frames) in training data into an optimal arrangement of bins to train the computational component which will ultimately be used to recognize new speech, after training is complete. The specification describes this as “. . . the identification of relevant acoustic contexts (i.e. phonetic contexts that produce significantly different acoustic feature vectors) is achieved through the construction of a binary decision network by means of an iterative split-and-merge procedure. The outcome of this bootstrap procedure is a domain independent general speech recognizer.” (Col. 7:41-47). As further elaborated in the ’925 Patent specification, a “split-and-merge procedure” allows the tree to be split (a node divided into multiple nodes) or merged (two nodes combined into one) as needed to maintain a good balance of how much of the training acoustic data goes to the right or left side at the node, depending on the yes or no answer to the question. (A decision tree is considered to be well balanced when the probabilities of the right or left branches for each node are approximately equal; a decision tree with well-balanced and distributed probabilities will tend to produce more accurate results than one where a high percentage of speech samples travel through just a concentrated subset of the nodes and branches). “A similar criterion is applied to merge nodes that represent only a small number of feature vectors” (col. 5:1-2). The decision tree can be used to sort the training data into “leaf nodes” (*i.e.,* the terminal nodes at the ends of the branches of the decision tree) which can then be used to train the acoustic portion of the recognizer. “The process stops if a predefined number of leaves is created. All phonetic contexts associated with a leaf cannot be distinguished by the sequence of phone questions that has been asked during the construction of the network, and thus are members

of the same equivalence class" (col. 5:6-10). So it can be seen that the decision network includes a mechanism to specify a set of phonetic contexts (based on the number of nodes produced in the split-and-merge process) and sort training data into collections for each of those contexts. This is one process for re-estimating the phonetic contexts where the nodes are not fixed. Once the training data is sorted to understand the resulting phonetic contexts, it can then be used to generate components that will later be used to actually recognize speech, including possibly Hidden Markov Models.

82. The corresponding limitation of '061 Patent claim 6 (according to Omilia's argument) recites "a set of states and a set of probability density functions assembling output probabilities for an observation of a speech frame in the states." This claim limitation thus includes states and probability density functions, where the probability density functions relate to outputs produced by the states, where those outputs are used in conjunction with ("observation") speech (frame) data. Although this is not limited to HMMs, HMMs are used in the specification to illustrate the claimed terms. "A Hidden Markov Model is a stochastic automaton… that operates on a finite set of states… and allows for the observation of an output each time… a state is occupied" ('061 Patent 3:18-21). This refers to the process of performing speech recognition. Recall that HMMs are used to classify input speech into its constituent phonemes by determining, on a frame by frame basis, which HMM would have been most likely to have produced acoustic features most similar to the speech. Further, note that this is a time-driven state model; each HMM represents one moment of time and the model is operated frame by frame (a stochastic automaton) on the input speech.

83. The claim additionally recites probability density functions. These PDFs quantify how likely it is that a given state would produce a particular set of acoustic features. "For any given HMM state, the unknown distribution of the feature vectors is approximated by a mixture of (usually Gaussian) elementary probability density functions (PDFs)." (Col. 3:37-39).

84. In short, the elements of claim 6 of the '061 Patent describe components which actively perform speech recognition, using a method in which states map to acoustic features

through particular probabilistic techniques, and the mapped features are compared to observed features of input speech, using the phonetic contexts that are available.

85. The decision tree disclosed in claim 1 of the ’925 Patent is a technique used in connection with generating a recognizer, sorting transcribed audio training data by recalculating a decision network that includes a tree-like structure automatically, using the phonetic context of the training data to determine how to cluster the data into distinctly recognizable sounds which are of phonetic significance. The states and PDFs from the supposedly corresponding claim 6 of the ’061 Patent are aspects of a computational technique to model the acoustics of different speech sounds in order to be able to identify those sounds later, while analyzing unknown speech input during recognition.

86. It is readily apparent that the two concepts in claim 1 and reference claim 6 are disjoint aspects of a speech recognizer. In the preferred embodiment, the ’925 Patent calls out components which are used to *produce* a recognizer through analysis recognizing and partitioning of phonemic contexts, both to identify a relevant set of such contexts, as well as to cluster training data for recognizing them. The ’061 Patent calls out a particular form of performing speech recognition—analyzing acoustic data—using well known state and probability based methods which can be used during the actual recognition process, by associating those states with acoustic features to be compared with the input features. The recognizer described in the ’925 Patent may typically also include HMMs similar to the ’061 Patent, but these are not called out as features which distinguish the claim. Similarly, a recognizer of the ’061 Patent *might* have used decision trees to initially determine which HMMs to instantiate in making up the set of states and PDFs, but that is not part of the invention as claimed.

87. Dr. Cohen argues that the claimed “set of states” of the ’061 Patent and their corresponding probability density functions correspond to the terminal “nodes” in the “decision network” claimed in the ’925 Patent. Cohen ¶ 42. Dr. Cohen argues the ’061 Patent at 4:16-19, 4:48-53 describes states as the terminal/leaf nodes of a binary decision network. However, these passages do not equate “states” with “nodes” of a decision network, and a POSITA would not

understand those terms to equate to each other. Rather, these passages illustrate how the decision network can be used to determine which phonetic contexts should be modelled (to adequately differentiate the samples in the training data), wherein each leaf node of the decision tree will be used to generate the related HMM. The leaf node represents part of a tree, used in one phase of operation, and the HMM represents a model which could produce acoustic data for that leaf node, in another phase of operation. Thus, a POSITA reading that passage would understand that the claims are directed to, on the one hand ('925 Patent), generating a recognizer by invoking the technique originally used to build the recognizer based on phonetic information, and on the other ('061 Patent), to simplify a recognizer, once already built, by simply removing portions based only on acoustic information.

88. This distinction is further borne out by the terms of the '061 Patent claim itself. There are states and probability density functions, which together can produce as output a speech frame (to be compared to the input speech frame). The states of HMMs have probability density functions (PDFs). PDFs describe a probabilistic output, such as "it's mostly like this, but could be a little like that, and is highly unlikely to be that other thing." Decision trees, however, are binary; a decision network is built so that each node in the decision tree asks a "yes"/"no" question about the observed acoustic frame, in order to determine a set of terminal nodes which will properly span the training set of acoustic frames. Decision tree nodes have no PDFs, and therefore, the decision tree nodes cannot be the claimed states.

89. As discussed above, the states of a model are operated in a discrete *time sequence* (hence the term, "state") to correspond to a series of input speech frames. This is why the '061 Patent refers to them as an "automaton" at col. 3:18. The nodes of the '925 Patent represent a *logical structure* which expresses how phonemes combine in a language (at least as evidenced by the training data). A "node" is well known in networking and graph theory as the junction of multiple links (the decision *network's* branches). A POSITA reading the specifications would not confuse "state" and "node" as referring to the same entity, and would not read the two claims as describing the same invention.

90. Dr. Cohen argues the ’061 Patent describes an acoustic model in terms of mathematical elements (“set of states” and “set of probability functions”) that comprise that model, while the ’925 Patent describes the phonetic structure of an acoustic model as abstract structures (“decision network with corresponding phonetic context”). Cohen ¶ 47. This argument seems to rest on the idea that the “mathematical” elements of the one and “structure” elements of the other are identical because they constitute alternate ways of expressing the same underlying model. Such a superficial analysis might lead to the conclusion that all speech recognizers are the same thing, in that they model some aspect of an underlying human language. But this fails to account for the detailed technical solutions espoused by each of the patents and their very different methods as set out in the two patents’ claims. Moreover, such a perspective does not preclude appreciating that the two patents describe distinct aspects of speech recognition—and the claim terms used in the ’061 Patent simply are distinct from the claim terms used in the ’925 Patent. To the extent they model aspects of human speech, the decision network models how phonemes combine into phonetic contexts for the language, while the states and PDFs model how those combinations might sound. One represents the cognitive aspects of how words are put together from underlying sounds, while the other represents the range of sounds the human vocal tract can produce to communicate those combinations.

91. At paragraph 48, Dr. Cohen is not offering a comparison of the claim language, but generally describes the state of the art of a “typical large vocabulary speech recognizer” that utilizes HMMs. What such recognizers may possess in common is beside the point of what the differences (described above, for example) are between what is specifically claimed by the actual language used in the reference claim and the ’925 Patent.

92. Similarly, at paragraph 49, Dr. Cohen’s argument generally about the use of an acoustic model to observe speech sounds also has no bearing on the differences between what is claimed in the two patents. The ’925 Patent describes a specific method of generating a second acoustic model from a first acoustic model that may start with a general purpose, large vocabulary speech recognizer, and by adding only a modest amount of additional specialized training data

("adaptation data," *see* col. 2:56) to the initial dataset, is able to efficiently provide an augmented speech recognizer by using the new training data to determine new phonetic contexts. The re-estimation process may result in new nodes of the decision tree, including terminal nodes, thus affecting how the acoustic model functions once it is deployed.

93. Further, the logic that a decision network and states with probabilities must be the same because they both rely on the same "acoustic model" is factually incorrect. As I said above, the decision network includes functionality that models how phonemes combine into phonemic contexts, *i.e.* the fundamental phonemic structure of the language and the words which constitute it. The states (such as HMM states) model the sounds of the language, *i.e.,* acoustic features which are produced by the vocal tract.

94. At paragraph 50, Dr. Cohen describes "nodes" of a decision tree and proposes that these correspond to states of HMMs. However, nodes are not a set of states. The decision network has a number of nodes, some *but only a minority* of which will be the resulting terminal nodes; these nodes represent the phonetic structure of the language. As discussed above, a decision tree will ultimately end in a series of terminal nodes (or "leaf nodes"), from which there is no further branching. While these *terminal* nodes (not all nodes) certainly *correspond* to HMMs (states and PDFs) useful for speech recognition, that does not make them the same, for all the reasons discussed above.

95. Paragraph 51 begins with a correct description of the squeezing process of the '061 Patent, but then the analysis misrepresents the role of the squeezed HMMs in order to claim similarity with decision networks. But there is no such similarity between squeezed HMMs and decision networks. Dr. Cohen opines that claim 6 claims a binary decision network, but this concept is never claimed in the '061 Patent. As I discussed in the previous paragraph, the presence of HMMs, which correspond to some of the terminal nodes of the decision tree, does not equate HMM states with decision tree nodes. Finally, Dr. Cohen states "These (PDF) functions represent the probability at each stage in the decision network and show the relationship between neighboring nodes…" This is factually incorrect on three counts. First, the PDF functions

associated with each HMM state represent probabilities that a particular set of acoustic feature values would be produced by the state, and has nothing to do with the decision network with corresponding phonetic contexts. Second, the decision network does not have probabilities associated with the nodes; they are binary. Third, the PDFs of one HMM are independent of any other HMM, and hence can bear no evidence to the relationship with any neighboring node.

96. At paragraph 52, Dr. Cohen relies on col. 4:48-51 of the ’061 Patent to conclude that the patent “defines” a set of states as context dependent word units, *i.e.* phones in context. He cites a passage from the ’061 Patent’s specification, which states, “the present invention suggests a reduction of the size of the acoustic model, i.e., both the number of acoustic subword unit HMMs, like, e.g. context dependent phone or triphone models, and the number of elementary PDFs for the output probabilities of these HMMs.” Col. 4:48-53. This passage from the specification is simply not definitional; it does not purport to define the meaning of “set of states.” At most, it only indicates that HMMs may be *built* for each phone in context. A phone in context is, for example, “the t between i and l as in “little”. An HMM can express what that “t” sounds like.

97. He further relies on col. 4:20-23 of the ’925 Patent to conclude that the set of states necessarily includes the phone context for each state. When he says “each terminal node (i.e. state)” he is expressing his opinion that nodes and states are the same, but I disagree for all the reasons stated above. The cited passage of the ’925 Patent provides “Since it is well known that speech sounds vary significantly with respect to different acoustic contexts, HMMs (or HMM states) usually represent context dependent acoustic sub-word units.” Col. 4:20-23. Remembering that an HMM is a mathematical construct used in speech recognition, and the pre-processing of the decision tree has identified which phonetic contexts are important enough to be modeled by an HMM, we can see that there is a clear *relationship* between the phonetic contexts classified by the decision tree and the HMM. In fact, that relationship is that the HMM represents what that phoneme in context sounds like. But it does not define the context as part of the phonemic structure of the language, as the decision tree does.

98. Finally, Dr. Cohen wrongly concludes that a POSITA would recognize that the "set of states" corresponds to a first decision network, but despite the correspondence between the states and the terminal nodes (only) of the decision network, this conclusion is fundamentally incorrect. The two aspects, set of states and decision network, are different components of the recognizer with different computational function. As such they represent different aspects of language (phonemic structure versus acoustic realization), and are used in different phases of recognition entirely (generating the recognizer from training data versus actually recognizing spoken speech), and employ terminology of the art (nodes in a network versus states in an automaton) which are fundamentally different operationally and conceptually.

2. "based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts" (green) [Cohen Decl. § VIII.C.4]

Claim 6 of '061 Patent	Claim 1 of '925 Patent
generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer . . . and generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer being distinctive of the particular application, such that the second speech recognizer is at least one tailored to the particular application and requires reduced resources compared to the first speech recognizer.	based on said first acoustic model, generating a second acoustic model with a second decision network and corresponding second phonetic contexts for said second speech recognizer by re-estimating said first decision network and said corresponding first phonetic contexts

99. The two claim limitations are directed to "generating" very different components of a speech recognizer. The cited limitation of the '061 Patent is directed to "generating … a set of states of the second speech recognizer" and "generating … a set of probability density functions of the second speech recognizer." By contrast, the cited limitation of the '925 Patent is directed

to "generating a second acoustic model with a second decision network and corresponding second phonetic contexts." These are very different. The first decision network and corresponding first phonetic contexts are not states. Phonetic contexts are groupings of phonemes in the context of their neighbors, and the decision tree determines the number of such groupings and how to sort the training data into clusters. States represent Hidden Markov Models, computational constructs to generate sounds with particular acoustic features, which features are compared to actual input speech during recognition. Dr. Cohen's opinion rests on the faulty premise that phonetic contexts can be equated with such acoustic models. *See, e.g.*, Cohen Decl. ¶¶ 58, 59.

100. Moreover, the methods of "generating" specified in the two patents are very different. The '061 Patent is directed to "selecting a subset" of states and PDFs from the first speech recognizer. By contrast, the '925 Patent is directed to "re-estimating said first decision network and said corresponding first phonetic contexts." Again, these are quite distinct. The fact that the '061 Patent requires a *reduction in resources* (per the claim language) for the second recognizer exemplifies the differences between claim 1 and claim 6; the '925 Patent re-estimates the decision network for the sake of improving recognition by enhancing performance or extending its scope, while the '061 Patent merely deletes other information (HMM states and corresponding PDFs) in order to shrink the recognizer while maintaining adequate performance.

101. Dr. Cohen mistakenly argues the re-estimation of claim 1 covers "adaptation" claimed by claim 6 which is the "selecting" element of claim 6. Cohen ¶ 58. The word "adaptation" does not even appear in claim 6. What does appear is "selecting a subset of states". It is clear from the specification of the '925 Patent that its claimed "re-estimation" involves recalculating the decision network in view of the new training data. This is so clearly different from "selecting a subset" that there is no comparison, which perhaps explains Dr. Cohen's reliance on the unclaimed concept of "adaptation" to complete his argument.

102. In what should take precedence over disagreement between myself and Dr. Cohen, the authors of the '925 Patent themselves explicitly declare that the two claim elements compared

here are completely distinct. The ’925 Patent states that “[t]his is considerably different from just “selecting” a subset of a decision network or “enhancing” the decision network:

> *Orthogonally to these previous approaches*, the present invention focuses on the re-estimation of phonetic contexts, or—in other words—the adaptation of the recognizer's sub-word inventory to a special domain. Whereas in any speaker adaptation algorithm, as well as in the above mentioned documents of *V. Fischer et al.,* the phonetic contexts once estimated by the training procedure are fixed, the present invention utilizes a small amount of upfront training data for the domain specific insertion, deletion, or adaptation of phones in their respective context. Thus re-estimation of the phonetic contexts refers to a (complete) recalculation of the decision network and its corresponding phonetic contexts based on the general speech recognizer decision network. *This is considerably different from just “selecting” a subset of the general speech recognizer decision network* and phonetic contexts *or simply “enhancing” the decision network by making a leaf node an interior node by attaching a new sub-tree with new leaf nodes and further phonetic contexts*. (Col. 6:66 - 7:17, emphasis added).

Note the use of the word “orthogonally”. To one as steeped in probability theory as the inventors, “orthogonal” means “completely independent” or “one having no bearing on the other.” The inventors were obviously aware of the techniques and limitations of each of their patents, and make it abundantly clear that they are not at all the same.

103. Thus, the ’925 Patent specification excludes processes that merely add nodes (attaching a new sub-tree), or selecting a subset, *i.e.,* deleting the nodes outside of the subset, without recalculating the decision network and its corresponding phonetic contexts. Thus even if

the HMM "states" of the '061 Patent corresponded to the "nodes" of the '925 Patent, the "generating" of each is quite distinct.

104. Dr. Cohen further argues that "re-estimation" of the '925 Patent may be accomplished by adding, deleting, or modifying phones in their respective context in the decision network. Cohen ¶ 41. I disagree. Dr. Cohen cites to a passage from the '925 Patent col. 7:2-12 out of context (Cohen ¶ 62), but it is clear when that passage is read in context of the entire paragraph that re-estimation refers to a "recalculation of the decision network and its corresponding phonetic contexts based on the general speech recognizer decision network." Thus, as outlined above, the '925 Patent distinguishes itself from the selecting procedures discussed in the '061 Patent. Although a re-estimation of the decision network and its corresponding phonetic contexts *may result* in the number of nodes changing, merely cutting down the number of nodes in itself does not constitute re-estimation. The two claims explicitly describe *distinct* methods of changing the recognizer.

105. Furthermore, in prosecution of the '061 Patent, the inventors made explicit statements that re-estimation and pruning are distinct, in successfully overcoming the Examiner's rejection based on Bahl. "Reestimating is <u>not synonymous</u> to pruning, as the Examiner suggests. Instead, reestimation is a separate and distinct function that can be performed to improve the recognition accuracy caused by pruning, since pruning typically creates a coarser approximation of the feature space for a given class…" **Exhibit B** ('061 Patent File History, 12/23/2003 Resp. to Office Action, underlining original).

106. Building on his misconception of what re-estimation means, Dr. Cohen concludes that because claim 6 results in a deletion of states (which he mistakenly claims correspond to nodes) the method is disclosed and included in the re-estimation process described in the '925 Patent. Cohen ¶ 61. I disagree. While the re-estimation process could result in terminal nodes being deleted, as well as other side effects, re-estimation is simply not adding or deleting nodes (either terminal or interior), but rather *generates* the recognizer, given the new adaptation data; "the limited, i.e. small, amount of adaptive (training) data suffices to generate the adapted speech

recognizer" (Col. 8:33-37). The method of the '061 Patent could be applied only later, *after* the re-estimation has been performed, as illustrated by '925 Patent claim 7. The '925 Patent describes a specific method of recognizing phonetic contexts that may have been under-weighted or even not have been originally present in the generalized speech recognizer, and then builds the resulting HMMs and determines their output probabilities, based on the new classification of feature vectors associated with leaf nodes. "The computation of an initial estimate for the state output probabilities …has to consider both the history of the context adaptation process and the acoustic feature vectors associated with each terminal node of the adapted networks." (Col. 8:44-48). In particular, note that the re-estimation takes into account both the original training data (which is carried forward in the phonetic contexts) and the new adaptation data. The '061 Patent can delete states and PDFs, but that is all it can do, and it can do these limited tasks based only on observing which HMMs actually contribute to recognition success given new adaptation data only.

107. In addition, the distinction between the two methods is explicitly stated in the '925 Patent when it points out that after the execution of the '925 Patent process, the '061 Patent process could *then* be applied. The '061 Patent process can be used *in addition to* and after the procedure of the '925 Patent, by re-using the adaptation data (which could be the same adaptation training data as in the '061 Patent) to squeeze HMM states. (Col. 8:58-9:3). This distinction by the inventors is further indication that the two techniques are disjoint. First, the decision tree models the *phonetic structure* of a language – which phonemes it uses, how they can be combined, and which ones sound similar based on which neighboring context. Importantly, in the process it does not simply list *every* phonetic context, but also groups them into ones that sound similar, perhaps even similar enough (the described equivalence classes) that there is no point in trying to distinguish them. Only then can the computational acoustic model of the HMMs be built, and, as stated, that model could be further squeezed.

108. Further, the fact that the number of nodes is not fixed in claim 1 would not lead a POSITA to conclude that any procedure that results in a changing of nodes satisfies the re-estimation procedure of the '925 Patent. Instead, is POSITA would conclude that the '925 Patent

claims a re-estimation process could recognize additional phonetic contexts, and could change what phonetic contexts are represented by what terminal nodes. The ’061 Patent cannot accomplish either of these because although the HMMs may correspond to individual phonetic contexts, the information to arrive at those contexts has been lost, and so can play no role in the process claimed by the ’061 Patent.

3. “based on domain-specific training data” (purple) [Cohen Decl. § VIII.C.5]

Claim 6 of ’061 Patent	Claim 1 of ’925 Patent
wherein selecting at least one of the subset of states and the subset of probability density functions of the first speech recognizer exploits application specific training data.	based on domain-specific training data,

109. Dr. Cohen argues that because the claims both use training data, the claim language is somehow “semantics and nothing more.” (Cohen Decl. ¶ 63). While both claim 6 and claim 1 use training data, the types of data are different and the claims use it to accomplish different things. In the ’925 Patent (col. 7:33-35) the data consists of acoustic frames, which are labeled (transcribed) either manually or automatically. As described in the ’925 Patent, training data is labeled, because the data is used to re-estimate the first decision network, for which each node asks yes/no questions based on knowing the preceding and following phonemes *(i.e.,* phonetic context). There is no need to consult the acoustic data to make such decisions, although the acoustic data will be clustered in the process. In the ’061 Patent (col. 6:12-15) the training data is just acoustic frames, which are associated with the particular HMM states by essentially performing recognition and keeping track of the states which produce the leading recognition candidates; this is what is referred to as “Viterbi aligning”.

110. In short, the data of the ’925 Patent is primarily phonetic labeling of speech frames (to be associated with decision tree nodes), while the data of the ’061 Patent is the actual acoustic contents of those frames (to be associated with the HMM states). Why are the labels in the ’925 Patent associated with frames at all? The answer is that the relative *numbers* (not contents) of the

frames help the tree calculation process (split and merge) determine how many nodes, both interior and leaf, are required to decide which questions should be asked to obtain a good spread of the data across leaf nodes. But, as the '925 Patent notes, as long as the acoustic frames are being used in this way, they can *also* be employed to perform the squeezing of the '061 Patent once this process is finished. The fact that this is a two-step process clarifies the distinction between how and which data is used in each patent.

111. As explained in the specification of the '925 Patent, after the method of the '925 Patent is performed, the '061 Patent may be further applied to optimize the speech recognizer.

> *Following the above mentioned teaching of V. Fischer et al., "Method and System for Generating Squeezed Acoustic Models for Specialized Speech Recognizer", European patent application EP 99116684.4*, **the adaptation data may also be used for a pruning of Gaussians in order to reduce memory footprints and CPU time**. The teaching of this reference with respect to selecting a subset of HMM states of the general purpose speech recognizer for use as a starting point ("Squeezing") and the teaching with respect to selecting a subset of probability-density-functions (PDFs) of the general purpose speech recognizer for use as a starting point ("Pruning"), both of which are distinctive of the specific domain, are incorporated herein by reference.

112. Indeed, this is precisely what is claimed in '925 Patent claim 7. The operations claimed in claim 7 correspond to the operations claimed in the '061 Patent. Accordingly, based on fundamental principles of claim construction, because the additional limitation claimed in dependent claim 7 covers the functionality described in the '061 Patent, that functionality *cannot* be what is covered by the limitations in independent claim 1 of the '925 Patent.

4. “wherein said first decision network and said second decision network utilize a phonetic decision [t]ree to perform speech recognition operations” (blue) [Cohen Decl. § VIII.C.6]

Claim 6 of ’061 Patent	Claim 1 of ’925 Patent
wherein the first speech recognizer includes a set of states and a set of probability density functions assembling output probabilities for an observation of a speech frame in the states,	wherein said first decision network and said second decision network utilize a phonetic decision [t]ree to perform speech recognition operations

113. Based on the claim language above, Dr. Cohen concludes that the recognizers from both patents use the “same” structure to recognize speech. (Cohen Decl. ¶¶ 41, 64-65 (citing ¶ 47-53)). I have already described at length the difference between “states” and “network” (as in phonetic decision network). *See, e.g., supra* ¶¶ 47-53. The network determines its leaf nodes, and these in turn cluster acoustic training data frames. From these acoustic frames, HMMs (the “states”) can be derived computationally. This does not make them the “same structure”. This basic logical fallacy is illustrated where Dr. Cohen states “The ‘set of states’ in the ’061 Patent is *from* a decision network that includes, by virtue of its structure, the relationship between a state and its neighboring states. Thus the set of states *is* a decision network…” (Cohen Decl. ¶65, emphasis added). Because the states are *derived from* a decision network does not imply that the states *are* the decision network. Paragraphs 66-70 merely reiterate this fallacy.

114. Dr. Cohen relies on the “bootstrap” procedure disclosed in the specification of the ’061 Patent in support of his argument that a POSITA would understand that a state is referring to a binary decision network. (Cohen Decl. ¶ 66). I disagree. The cited paragraph from the patent is clearly not calling out such an equivalence. Rather, it explicitly states that the binary decision network “separates the phonetic contexts into a pre-defined number of equivalence classes”. It is used for a process, of separating, and in the process gathers the acoustic feature frames into clusters. When the patent discusses a leaf node “defining” an HMM, this does not mean that the acoustic frames at the leaf node are already an HMM; instead one must be built based on that acoustic evidence. This bootstrap process is described in greater detail in the ’925 Patent (if the

two patents really do describe the same recognizer, then the bootstrap process would also be identical), in describing the leaf nodes: "Therefore the corresponding feature vectors are considered to be homogeneous and are associated with a context dependent, single state, continuous density HMM, who's output probability is described by a gaussian (sic) mixture model (eq. 4). Initial estimates for the mixture components are *obtained* by *clustering* the feature vectors at each terminal node, and finally the forward-backward algorithm known in the state of the art is used to *refine* the mixture component parameters." ('925 Patent col. 5:10-19, emphasis added). So, yes, the HMM "states" are related to the decision network leaf nodes, but by virtue of being derived from data at the leaf nodes. There is no equivalence.

115. Although Dr. Cohen describes similarities in structures of the recognizers in both patents in ¶ 67-68, this is actually revealing of distinctions between the two. Both patents include decision trees and HMM states. That the '925 Patent discusses and claims on the *network* while the '061 Patent discusses and claims on the *states* only shows to a POSITA that the inventions and claims are distinct. I have read Dr. Cohen's argument, that the states and networks are essentially the same thing, and a POSITA would not agree because the consistency in how the two patents refer to HMM states and decision trees or networks supports the conclusion that the terms refer to different things.

116. Furthermore, the last sentence of ¶ 67 makes a totally unfounded jump that the HMMs of the '061 Patent are organized into a *tree structure.* There is no evidence to support this, and HMMs are not in fact deployed in or representative of a tree structure. This is simple factual error. Indeed, as '925 Patent claim 6 expressly states, HMMs may be "associated with" leaf nodes, but are not themselves nodes in a decision tree.

117. A similar leap of logic is required by ¶ 68, that if the states of the '061 Patent are used to recognize speech (which is true) then the leaf nodes of the tree also must necessarily be used to recognize speech (which is false). The leaf nodes do not recognize speech. A POSITA at the time of the invention would know and understand this; thus, a POSITA would understand that the decision networks utilize the decision tree as discussed throughout my declaration to perform

speech recognition operations by determining the optimal way to cluster phonetic contexts, not that the leaf nodes are recognizing speech as Dr. Cohen mistakenly claims. HMMs derived from the acoustic features which are collected at the leaf nodes recognize speech.

118. The same error, about structure for derivation versus structure of the HMM states continues in ¶ 69. Again, the decision network contains nodes with corresponding phonetic contexts. The states, for which there may be one per *leaf* node, are not the nodes but rather derived from the nodes, and do not carry the structure of the decision network. Pointing out some similarities between the discussions in the two patents does not show that they in any way claim the same invention, or a subset of that invention.

119. Dr. Cohen's analysis relies on equating the "set of states" of the '061 Patent with the "decision tree" of the '925 Patent. For example, he opines that "[t]hese are different terms that refer to and cover the same subject matter." (Cohen Decl. ¶ 64; *see also* Cohen Decl. ¶ 47 (opining that the patents use "different words to describe these elements")); (Cohen Decl. ¶ 48 (opining that "the difference is one of semantics rather than the substance of what the claims cover")). But Dr. Cohen's attempt to re-interpret the two patents is misguided in light of the direct descriptions of both these components in both the patents.

120. The '061 Patent describes decision trees. It states that a decision tree is: "…a binary decision network that separates the phonetic contexts into a pre-defined number of equivalence classes" (col. 4:17-19). The '061 Patent also discusses states in the context of Hidden Markov Models ("HMMs") at length throughout, with a clear description: "A Hidden Markov Model is a stochastic automaton… that operates on a finite set of states… and allows the observation of an output each time… a state is occupied" (col. 3:18-22, mathematical notation omitted). There is no suggestion in the '061 Patent that "set of states" as used in claim 6 is defined to mean the same thing as "decision tree" as used in the patent's specification; accordingly, there is no basis to assume that the difference between "set of states" in the '061 Patent and "decision tree" in the '925 Patent is merely semantic.

121. Similarly, the ’925 Patent describes a “set of states” or subset of HMM states (col. 3:46, 8:64; claims 7, 20) and a “set of probability density functions” (col. 3:59, 6:64; claims 7, 20). This usage is fully consistent with the usage of those terms in the ’061 Patent. Similarly, there is no suggestion in the ’925 Patent that these terms are defined to mean the same thing as “decision tree.” Indeed, quite the opposite—they are consistently used either in portions of the specification that distinguish the invention of the ’925 Patent from the prior art, or in dependent claims whose very existence makes clear that these terms are different from “decision tree” as used in ’925 Patent claim 1.

122. Both patents are written in a technically precise style, complete with mathematical notation. The inventors of both patents were also accomplished authors of conference and journal papers, which require technical precision as well. The inventors took care to say precisely what they meant by these terms: “set of states” and “decision tree.” Both patents discuss both concepts, and they use the same term consistently in both patents, and also use the different terms in distinct manners. It is therefore inappropriate to attempt to “explain” how one of these concepts from one of the patents is really the same (except for “semantics”) as the other concept in the other patent. The inventors told us what they meant, and quite clearly.

5. “wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network,” (green) [Cohen Decl. § VIII.C.7]

Claim 6 of the ’061 Patent	Claim 1 of the ’925 Patent
generating, from the set of states of the first speech recognizer, a set of states of the second speech recognizer by selecting a subset of states of the first speech recognizer . . . generating, from the set of probability density functions of the first speech recognizer, a set of probability density functions of the second speech recognizer by selecting a subset of probability density functions of the first speech recognizer	wherein the number of nodes in the second decision network is not fixed by the number of nodes in the first decision network

123. Dr. Cohen previously asserted that the states of the '061 Patent are the same as the decision network of the '925 Patent, in analyzing the preamble to claim 1 (Cohen Decl. section VIII.C.2 in general). His analysis on this claim limitation relies on the same logic as earlier, and is similarly flawed in that the "states" and "decision network" describe different concepts and components of the recognizer, with the relationship between the two being limited to building an HMM for each *terminal* node *after* the entire process of recalculating the decision network is complete. Because Dr. Cohen reuses this claimed equivalence here, all my analysis to that previous claim element applies equally to this pair.

124. But the claimed equivalence is further distorted in this analysis, because here Dr. Cohen is equating the "states" not to the decision network as a whole, but rather to "nodes" within the network. This stretching of meaning further demonstrates the ill-fit of his states-equals-tree analysis. These claim limitations put further constraints on the nodes and states, respectively, derived from the processes by which they are processed by the claimed inventions.

125. Dr. Cohen argues that because nodes and states are purportedly the same thing, by claiming that the nodes in the second recognizer are not fixed by the nodes in the first decision network in the '925 Patent this covers states that are deleted in the '061 Patent, claim 6. I disagree with him for several reasons. First, nodes are not equal to states for the multiple reasons I have already discussed. (*See supra* ¶¶ 113-122). Second, the phonetic contexts and states of the '061 Patent are fixed by the phonetic contexts and states in the first recognizer. The second speech recognizer of claim 6 is "a *subset* of states *from* the first recognizer…" Thus, the second speech recognizer can only contain states found in the first speech recognizer. And if the states are phonetic contexts, as Dr. Cohen argues, the phonetic contexts in the '061 Patent cannot change; "…the phonetic contexts once estimated by the training procedure are fixed…" ('925 Patent col. 7:4-5, describing the priority document to the '061 Patent). The '925 Patent distinguishes itself from processes of the '061 Patent that merely delete nodes; "…the present invention utilizes a small amount of upfront training data for the domain specific insertion, deletion, or adaptation of phones in their respective context" ('925 Patent col. 7:5-8). This is because "the procedure uses a

maximum likelihood criterion to evaluate all possible splits of a node and stops if the thresholds do not allow a further creation of domain dependent nodes. This way one is able to derive a new, recalculated set of equivalence classes that can be considered by *construction* as a domain or dialect dependent refinement of the original phonetic contexts…" ('925 Patent col. 8:4-10).

126. Not only are the number of nodes not fixed, but in reality the decision tree is recalculated, not simply pruned as in the methods claimed in the'061 Patent. The '061 Patent does not recalculate the HMM states, it simply deletes some of them. Further, as described in the bootstrapping process of the '061 Patent (col. 4:11-27), the '061 Patent recognizer *already has its own decision tree*, which remains unchanged in the subsequent subsetting. It is inappropriate to equate states with decision trees when the '061 Patent has a component which much better matches the '925 Patent's decision trees, i.e., a decision tree. And there is no modification of that decision tree in the '061 Patent to correspond to the operations on the '925 Patent decision tree.

6. "and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer" (pink) [Cohen Decl. § VIII.C.8]

Claim 6 of the '061 Patent	Claim 1 of the '925 Patent
wherein selecting the subset of states comprises associating a multitude of speech frames of the training data with the correct states of the first speech recognizer	and wherein said re-estimating comprises partitioning said training data using said first decision network of said first speech recognizer

127. These claim elements point to core differences in the methods of the two patents. Dr. Cohen brushes aside the differences in these terms as "one of semantics." (Cohen Decl. ¶ 72). Calling the difference "semantics" reveals at least a lack of precision in understanding the two methods, even ignoring the differences between networks and states that I have discussed at length above.

128. In ¶73 Dr. Cohen claims that the "associating" of the '061 Patent "is the same process" as "partitioning" in the '925 Patent, but this is incorrect. The '061 Patent associates training frames with HMM states using the Viterbi algorithm. "For that purpose, usage of the

technique of Viterbi-aligning the speech data against its transcription to tag each speech frame with its correct leaf id." ('061 Patent col. 6:12-15). The Viterbi algorithm is the method used for recognition, and calculates for each HMM the probability that it produced the observed speech data. In doing so, it can be used to identify which series of HMMs match the input, or adaptation data. This is logical given the rest of the technique used by the '061 Patent, which will be to eliminate HMM states which are not found to be activated given the adaptation data.

129. This is not the same as the technique Dr. Cohen brushes past in ¶ 74. The '925 Patent explains "…each frame's feature vector is phonetically labelled and stored together with its phonetic context, which is defined by an arbitrary number of left and/or right neighboring phones. Subsequently the identification of relevant acoustic contexts (i.e. phonetic contexts that produce significantly different acoustic feature vectors) is achieved through the construction of a binary decision network by means of an iterative split-and-merge procedure." ('925 Patent col. 4:35-46). As discussed above, the split-and-merge example of re-estimation can separate nodes into subtrees, or merge nodes (both terminal and interior). And the '925 Patent goes on to describe this as asking a series of questions about the phones in the transcription and their positional relationships. Simply put, this is unrelated in any way to the Viterbi alignment on the HMM states described in the '061 Patent.

130. Finally, claim 6 goes on to read "and selecting those states with a frequency of occurrence above a threshold as the subset of states." This process is not used in the '925 Patent and there is no parallel operation. The first part of claim 6 recites determining which states are best fits (would "emit" the most similar sound) to each acoustic frame of speech training data, and this second part applies a simple threshold test to determine which states should be kept and which should be discarded. There is no simple threshold test described or implied in the '925 Patent. The re-estimation is as described above, an iterative split-and-merge procedure. There is nothing similar about these, even if the end results produced similar changes in leaf nodes and states (not that these are the same).

C. Step Two: Distinctions Between the Claims Render Them Patentably Distinct.

131. Because Dr. Cohen argues every claim limitation of claim of the ’925 Patent is the same as limitations found in claim 6 of the ’061 Patent, he does not perform analysis of the differences between the claims. However, the differences of the claims I have explained above, do make claim 1 patentably distinct from claim 6.

132. My analysis under Step One above has detailed differences between the ’061 and ’925 Patents, largely couched in rebuttal of Dr. Cohen’s contentions that the ’925 Patent claims a genus that includes within it the species claimed in the ’061 Patent. In this section I consider the inventive contribution of the ’925 Patent, specifically around the portions which Dr. Cohen incorrectly claimed were found in the ’061 Patent.

133. The primary contribution of the ’925 Patent is a specific method which allows efficient augmentation of a speech recognizer with adaptive training data from a particular domain to produce a new recognizer which is enhanced for that domain, but which also preserves much of its original capacity as a more general purpose recognizer. As was known in the art at the time, the first recognizer was built with extensive acoustic training data accompanied by phonemic transcriptions. The training data is too vast to be of use during recognition, so it must be reduced to meaningful subsets to represent as acoustic models. The subsets must be properly balanced, maintaining audible differences on the one hand, and adequately differentiating the phonemes in context on the other, while covering the full range of meaningful speech sounds. This is achieved by means of a phonetic context decision network, through processes that ask questions based on acoustic context in a manner which achieves optimal partitioning, and in the process recalculates the tree using split-and-merge algorithms, which can add or join nodes at any layer in the tree. Finally, the leaf nodes contain acoustic training data for each phonetically meaningful and acoustically distinguishable subset of the training data. The training data can then be modelled, where the models are used during the recognition process.

134. The phonetic context decision tree represents significant knowledge of the acoustic/phonemic structure of the language, being influenced by the frequency of actual contexts

encountered in training and their acoustic similarity. Essentially, it is a road map of what sound differences to pay attention to in order to understand the sounds of input speech as a series of phonemes.

135. A process such as described in the ’061 Patent has no access to that knowledge of the language, and can operate merely at the level of the sounds after they have been partitioned or sorted. And in fact all the ’061 Patent can do with those sounds (states) is delete some of them, by deciding they are no longer relevant based on additional domain specific data. This limitation is dictated by the fact that the ’061 Patent works only at the level of the acoustic data, and only with the purpose of making the recognizer smaller, in terms of computation and/or memory footprint.

136. The ’925 Patent is inventive in *maintaining* the original phonemic structure, so it can be *augmented* with additional domain specific training data, but without losing the rich context of the original phonemic decision tree. Essentially, the additional training data is used to tune the recognizer to perform especially well for a new domain, while still maintaining performance from its original, general, scope. The new, focused, training data is fed into the decision tree and is allowed to further modify it; the number of nodes in the tree is not fixed. This applies not only to the leaf nodes (which finally represent acoustic classes) but also to any sub-nodes and branches; the topology of the tree is allowed to change as a function of the new data. This process is referred to as “re-estimation”.

137. The advantage of this process is that the all the knowledge gained from training the original recognizer, as represented by its decision network, is fully available for augmentation in the derivation or re-estimation of the new decision tree. And as a result, the new decision tree is not limited in topology or number of nodes, carrying forward the extensive training of the original recognizer into the second recognizer. The process in the ’061 Patent is simplistic by comparison, only allowing for retroactively deleting the acoustic models, absence of knowledge of network topology.

138. The value of the process claimed in the ’925 Patent is demonstrated in the experiment summarized in the table at col. 9:40. The experiment compares the performance of a specialized digit-only recognizer, built by conventional means, with a “baseline” (general purpose) recognizer adapted for enhanced digit performance. The performance of these two recognizers is compared to the baseline recognizer, showing word error rates scaled to that of the baseline recognizer (as 100), for two tasks, a general dictation task and a digit recognition task. As can be seen in column 2 compared to column 1, the special purpose digit recognizer performs admirably on the digit task (error rate of 25 vs the baseline 100, a 4x improvement) but performs almost twice as poorly as the baseline recognizer on the dictation task. Looking instead at the adapted recognizer in column 3, it still performs twice as well as the baseline recognizer on the digit task, at a cost of only minor performance degradation (118) on the larger vocabulary dictation task.

139. It is precisely this ability to retain the original ability of the first recognizer in adapting it to a second domain which shows the value of the ’925 Patent invention. It does so in an efficient manner by not requiring the entire recognizer to be re-trained from scratch, because it retains and operates in the context of the full phonemic context derived for the original (general purpose) recognizer.

D. Comparing the Claims as a Whole.

140. In addition to all of the specific differences between individual claim limitations identified above, I have also considered the claims as a whole. In my opinion, as discussed above at length, the two claims are directed to entirely different solutions to entirely different problems. Moreover, Dr. Cohen’s analysis ignores key limitations of claim 6 of the ’061 Patent—all of the passages in his chart that are *not* color coded. These additional limitations are not mere surplusage, and are integral to the ’061 Patent’s specific solution to the problem it was aimed at addressing. There are at least four distinct limitations in ’061 Patent claim 6 that Dr. Cohen has ignored. First, “and requires reduced resources compared to the first speech recognizer” is fully consistent with and re-emphasizes the stated purpose of the ’061 Patent: to pare down a general purpose recognizer

for a more limited one that consumes fewer computational resources. Second, Dr. Cohen has entirely skipped the limitations of dependent claims 2 and 4, both of which further emphasize the particular function that claim 6 discloses, and that further illustrate how it differs from ’925 Patent claim 1. Finally, the last limitation of claim 6 requires “and selecting those states with a frequency of occurrence above a threshold as the subset of states.” This, too, emphasizes the function of cutting down the general purpose recognizer to eliminate states that are relatively unused by the specialized training data.

E. Claim 14 of the ’925 Patent Compared with Claim 15 of the ’061 Patent.

141. The substantive terms of ’061 Patent claim 15 (again, a dependent claim that includes the limitations of claims 10, 11, 13, 14 and 15) are the same as the substantive terms of ’061 Patent claim 6. Likewise, the substantive terms of ’925 Patent claim 14 are the same as the substantive terms of ’925 Patent claim 1. Therefore, my analysis of the differences between these claims is the same as my analysis above.

142. I further note, however, that ’061 Patent claim 15 is drawn to “an apparatus for automatically generating” while ’925 Patent claim 14 is drawn to “a computer-readable storage, having stored thereon a computer program having a plurality of code sections executable by a machine for causing the machine to automatically generat[e].” Dr. Cohen opines without analysis that these limitations in the preambles of the two claims “are equivalent.” (Cohen Decl. ¶ 75). Because that is a legal issue based on patent claim construction, and does not implicate any technical analysis of the differences between the claims, I do not address Dr. Cohen’s conclusion on that point.

F. Claim 27 of the ’925 Patent Compared with Claim 6 of the ’061 Patent.

143. Claim 27 contains similar claim limitations to claim 1. Claim 27 further adds limitations concerning a first and second language that are the domains and the second speech recognizer is multilingual: “wherein the first domain comprises at least a first language, wherein

the second domain comprises at least a second language, and wherein the second speech recognizer is a multi-lingual speech recognizer."

144. Dr. Cohen makes three arguments concerning claim 27: (i) incorporating his opinions concerning claim 1, (ii) that "domain-specific training data" in claim 27 is the same as "application specific training data" are the same, and (iii) that the final limitation of claim 27, "wherein the first domain comprises at least a first language, wherein the second domain comprises at least a second language, and wherein the second speech recognizer is a multi-lingual speech recognizer," though not actually present in claim 6 of the '061 Patent, would have been obvious to a POSITA in view of the cited prior art.

145. As to the first argument, it is my opinion that claim 27 is different from claim 6 for reasons identified above with claim 1. As to his second point, I refer to and incorporate my comments, above, concerning how the training data is used (*see supra* ¶¶ 109-112). Finally, in my opinion, the final limitation of claim 27, "wherein the first domain comprises at least a first language, wherein the second domain comprises at least a second language, and wherein the second speech recognizer is a multi-lingual speech recognizer," is simply absent from claim 6 of the '061 Patent. Therefore, this limitation, and claim 27 as a whole, are not an obvious variant of '061 Patent claim 6. Dr. Cohen's extensive reliance on additional purported prior art references does not change my conclusion that claim 27 is not an obvious variant of claim 6. Instead, Dr. Cohen's analysis appears to be more akin to a traditional obviousness analysis; however, since ODP is limited to the claims of the earlier patent, it is not permissible to combine the disclosures of those claims with other alleged prior art.

146. Because his argument ignores the differences I have set out above, his argument is misguided and irrelevant. For example, Dr. Cohen argues that because the first speech recognizer can include a first and second language based on the Court's claim construction, the multilingual limitation does nothing more than apply claim 6 to a pre-existing multilingual speech recognizer and multilingual speech recognizers were known in the art. (Cohen Decl. ¶ 82-84). Not only is this conclusory argument lacking in any explanation as to why this would be the case, it is the

wrong inquiry because claim 27 claims a completely different method for generating and adapting a speech recognizer as explained above. Based on the invention of the ’925 Patent, a multilingual speech recognizer can be created, e.g. through an iterative process to enhance the general recognizer's phonetic contexts incrementally based upon further training data, where the domain specific data contains the second language. The ’925 Patent accounts for phonetic contexts available in the training data that are not already part of the first speech recognizer. Reference claim 6 does not.

147. Additionally, Dr. Cohen has impermissibly relied on the ’061 Patent specification throughout his argument in this section as prior art. I understand that as part of an ODP analysis, the specification of the reference patent cannot be used as prior art. But this is precisely what Dr. Cohen has done. He has looked to the specification and argued that because the ’061 Patent can be used to adapt a recognizer to dialects, then it can be used to adapt a recognizer to be multilingual (Cohen Decl. ¶ 84). Dialects are not covered in claim 6 and should not be considered in this analysis—it not the specification as a whole but what the patent *claims* that is relevant. Because Dr. Cohen is relying on the ’061 Patent specification as prior art to conclude a POSITA would find claim 27 obvious in view of other purported prior art, his entire analysis suffers from this flaw and should not be considered.

148. Even though Dr. Cohen’s argument is based on faulty premises, I nonetheless have addressed some of his arguments because they do not make sense at face value. The ’061 Patent method would contain only a subset of states and cannot account for phonetic contexts found in the new training data of a second language. In discussing the adaptation of its recognizer for another dialect, *e.g.* Austrian German, the ’061 Patent does not add any HMM states. The description beginning at col. 8:36, and in particular the equation in the midst of this description, shows how the process operates. As before, additional training data, this time from the new dialect, is fed through the recognizer. Instead of simply determining which components of the Gaussian Mixture Model (PDF components) are under-utilized, and hence subject to deletion (pruning), the equation shows that the decision is biased in favor of the new training data. Thus the new training

data, *i.e.* the data corresponding to the desired new application, is given more importance than the generic training data when making the decision as to whether a particular Gaussian should be squeezed.

149. The adaptation to a specific dialect in the ’061 Patent thus consists of simply removing those PDF components of the general recognizer which are not found in the training data from the dialect. If there are some sounds or ways of speaking in the general language recognizer that do not occur in the dialect, then they cannot help recognition performance in the specialized (dialect) recognizer, and can be removed, to achieve the goal of a smaller computational or memory footprint, in the specialized recognizer.

150. By comparison, in the ’925 Patent, according to claim 27, the phonetic context and decision network are modified during the process of recalculating the first recognizer, which includes at least one language, to a second recognizer that includes at least a second language, by adding domain-specific training data from the second language. The second language will share many speech sounds with the first, but may well have some sounds not found in the first language. This new training data will change the decision network, but all the training which went into the first recognizer carries over into the second, compound recognizer. The newly acquired training data from the second language would then be used to generate the HMMs associated with newly created leaf nodes.

151. The two approaches are quite distinct. The ’061 Patent operates at the level of the HMMs requiring states and Gaussian Mixture Models (GMMs) requiring probability density functions (PDFs), and can only delete pronunciations (states and PDFs). The ’925 Patent operates at the level of the decision network/phonemic context– the phonemic mix of both languages, when during adaptation, it mixes in the phonemes of a new language.

152. Paragraph 85 is equally misguided because it relies on conflating the methods of claim 6 with the methods of claim 27, without considering any of the differences in the two methods. The issue is not that multilingual speech recognizers existed, but that claim 6 and claim 27 claim entirely different ways of adapting speech recognizers. Dr. Cohen simply mentions

speaker dependent recognizers and other art and several papers about multi-lingual recognizers, neither fact being particularly relevant to the differences between reference claim 6 of the ’061 Patent and claim 27 of the ’925 Patent. The statement implying that building a speaker dependent recognizer and a multilingual recognizer using processes which vary only on the basis of training data is factually incorrect and unsupported. The ’925 Patent disagrees at the discussion of MAP and MLRR at col. 6:36-43.

153. Dr. Cohen refers to the Schultz paper attached to his declaration as Exhibit D, “Multilingual and Crosslingual Speech Recognition,” apparently to demonstrate that multilingual recognizers were known in the art. I do not dispute this general observation, but fail to see how Dr. Cohen has demonstrated that any concept in this publication mapped to the invention disclosed in the ’925 Patent. Further, in conducting the ODP analysis, I have been informed that it is not permissible to combine the reference patent with other prior art. In fact, the paper discusses multilingual recognizers which are designed up front for multiple languages, based on creating an entirely new recognizer by initially combining training data from multiple languages which have been pre-selected to give broad phoneme coverage. This is different from the ’925 Patent process of re-estimation to augment a first recognizer (that includes a first language) with a second domain that ultimately results in a multi-lingual recognizer. The paper also discusses “crosslingual” recognition in which “the developed multilingual systems are applied to recognize *new unseen* languages without any additional training” (Schultz Exhibit D at p. 5, emphasis original). The ’925 Patent is clearly a different technique which relies on additional training data.

154. Dr. Cohen then offers a hypothetical example of a recognizer multilingual in three languages, being given additional training data in two of those languages, with a resulting recognizer for some reason recognizing only two languages. Perhaps this is meant to dovetail with the fact that the ’061 Patent can only delete or remove recognition ability, and so further constrain the recognizer. But the given example, which is apparently to show how the ’061 Patent could generate a multilingual recognizer by starting with a multilingual recognizer, simply would not work as described. In Dr. Cohen’s example, the original recognizer knows Portuguese, Italian,

and Spanish phonemes. It is fed adaptive data for Italian and Spanish. According to the ’061 Patent process, sounds not heard in the adaptation data (*i.e.,* the Portuguese phonemes) will have their associated states *deleted.* So the resulting recognizer could not, as Dr. Cohen asserts, recognize Portuguese (as a “first language”) and Italian (as a “second language”) because all the Portuguese-specific sound models would have been deleted.

155. Further, creating a multi-lingual recognizer by starting with a multi-lingual recognizer and removing the ability to recognize one of the languages does not analyze the differences between the claims in determining whether claim 27 is patentably distinct. If anything it underscores why a POSITA would not apply claim 6 to developing a multilingual speech recognizer. First, it would be strange indeed if a POSITA, wanting to generate a recognizer for some pair (or more) of languages would *start* with a recognizer which could already recognize those languages, and then some, and look to delete the language(s) not needed. One wouldn’t even consider the ’061 Patent because it is directed to *shrinking* a recognizer (in order to employ it in computational constrained situation) rather than *extending* the recognizer to add the capability of recognizing a second language. Second, this points out just why the ’925 Patent does not reflect the claims of the ’061 Patent at all. The ’925 Patent strives to maintain, and extend, phonetic richness, in the form of the phonetic decision tree, by augmenting and re-estimating it based on new adaptation data (containing, in this case, the sounds of the second language which are not found in the first language). Of course, this is almost certainly going to use more computational resources; this is the exact opposite of the requirement placed on the ’061 Patent.

156. Paragraph 86 adds nothing to the analysis as it simply mentions a publication (Schultz, Cohen Decl. Exhibit E), which was cited on the face of the ’925 Patent; of course multi-lingual recognizers were known at the times of each of the patents in question. Dr. Cohen has not analyzed how these speech recognizers were created using steps of claim 27 that are different from claim 6. The Examiner cited this publication and still found claim 27 patentable (**Exhibit C** (’925 Patent file history, Office Action dated 18 May 2005)); the applicants overcame this rejection simply by moving some limitations of dependent claims into the independent claims. As with the

Schultz paper cited as Cohen Exhibit D, this paper also attempts to build a global phoneme pool by combining phonemes from a number of languages, and then specializing this collection for a particular language, a different technique from that of adapting a given recognizer for a new language as claimed in the ’925 Patent. Further, in conducting the ODP analysis, I have been informed that it is not permissible to combine the reference patent with other prior art. It is clear that Dr. Cohen’s analysis is *not* simply using this prior art in order to construe claim language as part of his step one analysis, or to show that a POSITA would understand some limitation of ’061 Patent claim 6 to correspond to this limitation. There is no corresponding limitation in claim 6, and Dr. Cohen is using the prior art either in combination with the ’061 Patent, or instead to completely read this limitation out of ’925 Patent claim 27. Neither is permissible.

157. Dr. Cohen further argues that the methods of claim 6 can be used to expand a single language recognizer to a multi-lingual recognizer at paragraph 87. First, this is not the appropriate inquiry here as Dr. Cohen has supplanted claim 27 for claim 6 of the reference patent, completely ignoring the differences. Second, Paragraph 87 seems to somehow confuse the explicit mention of HMMs in claim 6 of the ’061 Patent with multi-lingual recognition, simply because HMMs model speech sounds and a language is composed of speech sounds. The ’061 Patent would adapt to a dialect (again, dialects are not claimed in the ’061 Patent and should not be considered in this analysis) by deleting HMMs corresponding to sounds found in the general language recognizer (*e.g.* German) which are not found in a particular dialect of that language (*e.g.* Austrian German). The remainder of the paragraph makes no sense, unless the expert does not appreciate that Austrian is not a distinct language, but rather simply a dialect of the German language. Dr. Cohen fails to show how removing speech sounds from a single-language recognizer can turn it into a multi-lingual recognizer. Nor does he give any indication of how removing pronunciations could possibly “apply equally” to adding the phonemes of a new language to a recognizer, and most significantly, how any of this applies to the techniques claimed in the ’061 Patent.

158. Again, in Paragraph 88 Dr. Cohen impermissibly seeks to rely on the specification of the ’061 Patent as prior art, but it seems to be an attempt by Omilia’s expert to re-write the ’061

Patent in a way not disclosed in any form in its specification—neither of which are permissible or make sense. Specifically, Dr. Cohen states that, "one way to do this would be to add nodes for phones in training data" but this is Omilia's expert making up his own invention. Since the '061 Patent never discusses changing nodes in a decision tree, but only using it for bootstrapping the original recognizer, the expert's opinion that such could have been done is not suggested in any way by the specification and is therefore irrelevant. Nothing in the '061 Patent begins to suggest that those resources might be used by adding new HMMs corresponding to new phonemic contexts associated with a new language. In fact, since the '061 Patent operates only at the level of the HMMs and PDFs, having used a phonemic decision tree only to generate these in the bootstrapping process for the base recognizer, it would be quite at odds with the specification to do so.

159. The premise of the '061 Patent is that one can start with a general purpose recognizer and configure it to perform adequately for a particular application by observing the speech sounds found to be most relevant to that application, and removing portions of the recognizer which contribute only to recognizing other speech sounds, *i.e.* removing HMM states and/or associated output PDFs. In the process of doing such, a developer would run training data through the system, detect unneeded states or PDFs, remove them, and verify that recognition performance was not unduly compromised. One might indeed find, as the '061 Patent points out, that acceptable performance can be obtained using even fewer computational or memory resources. Indeed, then the excess saved resources could be used in some other aspect of the recognition process to improve performance further. But absolutely nothing in the '061 Patent suggests that those "spare cycles" be used for the totally unrelated function of possibly adding terminal nodes by re-estimating a decision tree. Yet a new language requires new nodes in the tree, as it has different phonemic contexts, and almost certainly even new base phonemes than the original language – a very different proposition from removing states to optimize recognition on a dialect. Spare resources could easily be employed in language modeling, for example (a part of any large vocabulary recognizer, but beyond the scope of either patent); decision trees are mentioned only in the initial, bootstrapping, phase for the '061 Patent recognizer and there is no

suggestion of performing the additional operations on them required for adapting to a new language.

160. Beginning at paragraph 89, Omilia's expert opines on a process for building a multi-lingual recognizer as disclosed in the Sabourin patent, which is completely distinct from claim 27 of the '925 Patent.

161. Dr. Cohen argues that a POSITA would have combined the teachings of Sabourin with the '061 Patent "to account for phones in the second language or domain by adding a node to recognize phones not present in the first speech recognizer and generating a multilingual speech recognizer as taught by Sabourin." I disagree for a number of reasons.

162. Foremost, Dr. Cohen's opinion does not consider the difference between reference claim 6 and claim 27, but relies on the specification of the '061 Patent. This is improper under this analysis. I understand that for obviousness-type double patenting analysis, it is the claims, not the specification, that define an invention and it is the claims that are compared when assessing double patenting. Dr. Cohen does not even consider the differences in claims of the '925 Patent at all in his analysis. This is telling because Sabourin does not contribute in any way to the claims of the '061 Patent in comparing them to the '925 Patent.

163. Further, in conducting the ODP analysis, I have been informed that it is not permissible to combine the reference patent with other prior art. Thus, Dr. Cohen's reliance on Sabourin is improper. Dr. Cohen makes no substantive analysis of Sabourin, simply making vague suggestions that Sabourin somehow suggests how claim 6 can be used to accomplish the same end as claim 27, without even touching on which aspects of Sabourin would support such a suggestion or, more importantly, how Sabourin harmonizes with claim 6 in a manner which shows it is not distinct from claim 27. In this analysis Dr. Cohen suggests that Sabourin discloses "the method of the '061 Patent with the additions of nodes to generate a multi-lingual recognizer." (Cohen ¶89) But the '061 Patent can only delete states (which Dr. Cohen equates with nodes) so how could the opposite teaching be suggested? It certainly is not found in Sabourin.

164. Sabourin describes in detail a process for creation of a multi-lingual recognizer. It begins with letter to sound rules (found in neither the ’061 Patent nor the ’925 Patent, both of which use phonetically labelled or HMM aligned training data) which allow training data to be assigned to an apparently hand selected set of pre-determined phonemes for the joint languages. This rule set is not built computationally but rather by a trained linguist (col. 5:21-22). Phonemes in the second language are initially mapped to the nearest phonemes in the first language, not by a decision tree, but rather by evaluating similarity, based on a phonemic feature description of the phonemes (again, determined by a linguist or phonetician), finding which old phoneme is minimally different from the new one by minimizing a “transformation” rule, which basically compares which articulatory features are similar or different across the set of old phonemes.

165. After this is done, the recognizer for the new language can be further adapted to the new pronunciations using maximum a posterior adaptation (MAP), which is a process of updating HMM models with new data after they have been initially modeled by training data. But this technique is specifically taught away from in the ’925 Patent and the ’061 Patent. “Adaptation techniques known the (sic) within the state of the art, for example maximum a posteriori adaptation (MAP) or maximum likelihood linear regression (MLLR)… exclusively target the adaptation of the HMM parameters based on training data. Importantly, these approaches leave the phonetic contexts unchanged; that is, the decision network and the corresponding phonetic contexts are not modified by these technologies.” (’925 Patent at col. 6:36-46); (*see also* ’061 Patent at col. 4:29-34 (“The proposed solution according to the present invention to the above-mentioned problems is orthogonal to approaches which exploit speaker adaptation techniques, like, e.g., maximum a posteriori adaptation (MAP)…”).

166. Eventually Sabourin does mention decision trees, but only in the context of splitting phonemes into allophones – *i.e.* variants of the phoneme with different acoustic realizations. But unlike the ’925 Patent, which could possibly end up with allophones in different leaf nodes, a separate HMM is not trained for each, but rather the allophone is adapted from the original HMM using, once again, MAP adaptation (col. 11:34). So the only similarities between Sabourin and

the ’925 Patent are that it deals with multiple languages and makes temporary use of a decision tree in a limited context.

167. Thus Sabourin does not resolve the differences between the ’061 and ’925 Patents, or even contribute relevant teachings or employ related methods. When properly considering the differences discussed above, claim 27 is not an obvious variant of claim 6. Removing HMM states and PDFs is no more similar to re-estimating decision trees in the context of multi-lingual recognition than it was in the analysis of claim 6 against claim 1. The emphasis on removing rather than adding or enhancing recognizer footprint simply does not speak to accounting for phonetic contexts of the second language.

168. Second, even if the specification of the ’061 Patent were considered in this analysis, this would not change my opinion because nothing in the specification of the ’061 Patent would render claim 27 obvious in view of Sabourin. The ’061 Patent is such a different technique, that it would have no place in operating alongside or in addition to Sabourin, and Dr. Cohen certainly doesn’t present such an analysis.

169. Next Dr. Cohen argues that a POSITA would have combined Sabourin and the ’061 Patent to generate a multilingual speech recognizer. This is beside the point. The appropriate test here requires that the differences between the claims be considered in view of the prior art, and Dr. Cohen has not performed this analysis. Whether the ’061 Patent could be applied to a pre-existing multilingual speech recognizer does not obviate the method of claim 27 by which a multilingual speech recognizer is adapted. Taking Dr. Cohen’s argument at face value (that claim 6 of the ’061 Patent can be performed on a pre-existing speech recognizer that remains multilingual), this would not lead a POSITA to conclude that claim 27 is obvious variant of claim 6 for the same reasons I have already cited above.

170. Finally, Dr. Cohen cites to Schultz in his opinion at paragraph 93 to conclude the same thing he does with Sabourin, that a POSITA would combine the ’061 Patent claim 6 with another Schultz paper (Exhibit G to his declaration) to create a multilingual speech recognizer. First, in conducting the ODP analysis, I have been informed that it is not permissible to combine

the reference patent with other prior art, much less to combine multiple pieces of prior art, as Dr. Cohen does here. But further, he refers to Schultz as taking a German speech recognizer and adapted it to the Japanese language by copying the codebooks of the German recognizer for similar sounds in the Japanese language. (Schultz at 3). When a Japanese sound did not have a German equivalent, the recognizer took the most similar sound from German and used its codebook. (Schultz at 3). It is not evident how this doubling of some phonemes is in any way related to the deletion of phonemes (or states) in the '061 Patent. Further, the doubling in Schultz is contradictory to the methods of the '925 Patent, which maintain the richness of the original decision network while enhancing, or adding to it, with the new adaptive training data. There is no need to "make do" with existing contexts in the '925 Patent, because its very goal is the ability to incorporate new contexts into the decision tree generation of the multilingual recognizer.

171. When the appropriate inquiry is made into the differences of the claims and the prior art, such as Sabourin, a POSITA would conclude that the '925 Patent is patentably distinct. None of the prior art suggests generating a new recognizer by starting with an existing decision network and incorporating new adaptation / training data and running a re-estimation process, for the sake of deriving a desirable set of distinctive and well weighted phonetic contexts to form the basis of acoustic models thereof. And certainly the teachings of the '061 Patent, which are limited to the removal of states with the explicit goal of simplifying the recognizer, do not lend credence to such an analysis.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct.

Executed on January 6, 2021 at Winchester, MA.

Chris Schmandt

CERTIFICATE OF SERVICE

I hereby certify that this document and its exhibits and attachments thereto will be filed through the ECF system and will be sent electronically to the registered participants as identified on the Notice of Electronic Filing (NEF) and paper copies will be sent to those indicated as non-registered participants on January 7, 2021.

/s/ Christian E. Mammen
Christian E. Mammen

Christian E. Mammen (*admitted pro hac vice*)
WOMBLE BOND DICKINSON (US) LLP
1841 Page Mill Road, Suite 200
Palo Alto, CA 94304
Telephone: (408) 341-3067
Chris.Mammen@wbd-us.com